CLINICAL STUDY

Validity and reliability of the novel thyroid-specific quality of life questionnaire, ThyPRO

Torquil Watt^{1,2}, Laszlo Hegedüs³, Mogens Groenvold^{2,4}, Jakob Bue Bjorner^{2,5}, Åse Krogh Rasmussen¹, Steen Joop Bonnema³ and Ulla Feldt-Rasmussen¹

¹Department of Endocrinology, Copenhagen University Hospital Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark, ²Institute of Public Health, University of Copenhagen, DK-1014 Copenhagen, Denmark, ³Department of Endocrinology and Metabolism, Odense University Hospital, Odense, Denmark, ⁴Department of Palliative Medicine, Bispebjerg Hospital, Copenhagen, Denmark and ⁵National Research Centre for the Working Environment, Copenhagen, Denmark

(Correspondence should be addressed to T Watt at Department of Endocrinology, Copenhagen University Hospital Rigshospitalet; Email: t.watt@rh.dk)

Abstract

Background: Appropriate scale validity and internal consistency reliability have recently been documented for the new thyroid-specific quality of life (QoL) patient-reported outcome (PRO) measure for benign thyroid disorders, the ThyPRO. However, before clinical use, clinical validity and test–retest reliability should be evaluated.

Aim: To investigate clinical ('known-groups') validity and test–retest reliability of the Danish version of the ThyPRO.

Methods: For each of the 13 ThyPRO scales, we defined groups expected to have high versus low scores ('known-groups'). The clinical validity (known-groups validity) was evaluated by whether the ThyPRO scales could detect expected differences in a cross-sectional study of 907 thyroid patients. Test–retest reliability was evaluated by intra-class correlations of two responses to the ThyPRO 2 weeks apart in a subsample of 87 stable patients.

Results: On all 13 ThyPRO scales, we found substantial and significant differences between the groups expected to have high versus low scores. Test–retest reliability was above 0.70 (range 0.77–0.89) for all scales, which is usually considered necessary for comparisons among patient groups, but below 0.90, which is the usual threshold for use in individual patients.

Conclusion: We found support for the clinical validity of the new thyroid-specific QoL questionnaire, ThyPRO, and evidence of good test–retest reliability. The questionnaire is now ready for use in clinical studies of patients with thyroid diseases.

European Journal of Endocrinology 162 161–167

Introduction

Measurements applying standardized self-reports to capture the impact of health on patients' lives are termed health-related quality of life (HRQL) measurements (1). They usually conceptualize HRQL as a multidimensional concept encompassing various aspects of physical, mental, and social functioning and well-being. To an increasing extent, the broader, but also more neutral term 'patient-reported outcomes (PROs)' is replacing HRQL. Today, PROs or HRQL measurements are recognized as inevitable and important outcomes in high quality clinical studies. Further, they can provide important documentation for evidence-based patient information and may even be implemented in clinical management of the individual patient, as has been done within, e.g. oncology (2), where randomized trials have documented significant

improvement of patient–clinician interaction, without prolonging consultations, and impact on patient management (3). HRQL measurements may be either generic, i.e. applicable to any patient group regardless of diagnosis, or specific, i.e. targeted to a specific disease group. Specific HRQL measurements are usually more sensitive than generic, which on the other hand have the advantage of allowing comparisons across dissimilar populations.

Some questionnaires have been developed for specific thyroid diseases (4–12). However, a thoroughly validated questionnaire only exists for thyroid-associated ophthalmopathy (TAO) patients (5–7). Another TAO questionnaire has been developed, but has not been validated (8). One questionnaire for patients with hyperthyroidism was developed, but has never been validated (4). Three questionnaires for hypothyroid patients have been developed (9–12), but studies

evaluating the validity of these measures are still awaited. Most importantly, no validated, thyroid-specific PRO instrument is available for use across different thyroid diseases (13). This is a major deficiency, because benign thyroid diseases are characterized by a substantial overlap between various disease entities (e.g. coexistence of goitre and hyperthyroidism) and a shift between diseases (e.g. hyperthyroid patients becoming hypothyroid through ablative therapy). Therefore, optimally an HRQL outcome measure for thyroid patients should encompass all thyroid diseases in order to have content validity (i.e. capture HRQL issues of relevance to the patients). If not, the results of longitudinal studies may be misleading, because important HRQL aspects are not measured at follow-up (e.g. impact of hypothyroidism after ablative therapy).

We have recently developed a quality of life (QoL) questionnaire for patients with benign thyroid diseases, called the ThyPRO (13–15), and evaluated important aspects of its measurement properties (16). In HRQL terminology, the measurement property termed accuracy within biochemical assay methodology (i.e. degree of systematic bias) is called validity and what endocrinologists may refer to as precision (or reproducibility) is termed reliability. In the just mentioned validation study (16), support for appropriate validity and reliability was found, in terms of a valid scale structure (an important aspect of what is termed construct validity) and very good internal consistency reliability. However, for HRQL measurements, evaluation of accuracy, or validity, is a more complex task than with most other measurement fields, since no gold standard against which other measures can be tested is available. Thus, the validation of an HRQL measure is an iterative process where evidence for or against the validity of a measure is usually gathered by several studies taking different approaches to evaluating validity, one of which is the above mentioned scale validation. Another important way of assessing validity, especially for measures attempted for clinical use, is termed known-groups validity. In this approach, clinically based criteria are used to classify patients into groups with expected high or low scores on a questionnaire and then test whether these expected differences are found in patient samples. Precision, or reliability, can also be evaluated by several techniques. One approach is 'internal consistency reliability' (Cronbach's alpha), which has been used in the initial analyses of the ThyPRO (16). Another approach is test–retest reproducibility where duplicate measurements are obtained by collecting two responses from stable respondents separated by 2–3 weeks (17).

The purpose of the present study was to investigate accuracy, in terms of clinical/known-groups validity, and precision, in terms of test–retest reliability, of the Danish version of the thyroid-specific QoL questionnaire, the ThyPRO.

Material and methods

Patients and clinical characterization

Patients were recruited from the endocrinological outpatient clinics at two university hospitals in Denmark: Copenhagen University Hospital Rigshospitalet (RH) and Odense University Hospital (OUH). At RH, the sampling strategy was cross-sectional: all thyroid patients born within the first 20 days of each month (to limit running sample size) were invited during February–June 2007 by mail 3 weeks prior to their appointment in the clinic. At OUH, all eligible patients referred to the thyroid unit of the endocrine outpatient clinic during May–November 2007 were recruited. The questionnaire was sent about 3 weeks prior to the appointment in the clinic. Blood samples were drawn the week prior to their appointment, and the participants were instructed to complete the questionnaire at about that time. Questionnaires were either returned by mail or delivered by hand in the laboratory or at the clinic on the day of appointment. Exclusion criteria were absence of any thyroid disorder, thyroid cancer, age <15, and inability to complete a questionnaire due to communication problems (non-Danish speaking, blindness, etc). One reminder was sent after 2 weeks to non-responders, and all participants gave signed informed consent.

Socio-demographic data and information about co-morbidity and non-thyroid medication were self-reported. Laboratory data, diagnostic imaging results, exact diagnosis, previous and current treatment and time of diagnosis among respondents were obtained by chart review. Biochemical thyroid tests were TSH, total thyroxine (T₄), total triiodothyronine (T₃), free T₄ (FT₄), free T₃ (FT₃), resin-T₃ test (only OUH), thyroid peroxidase antibodies (TPOAb), thyroglobulin antibodies (only in patients diagnosed with hypothyroidism and negative TPOAb) and TSH receptor antibodies (only in patients with a diagnosis of hyperthyroidism). All analyses were performed using the standard methods at the laboratories of the participating hospitals. In patients with thyroid eye disease TAO, NOSPECS (18) classification was performed by an ophthalmologist and clinical activity scoring (19) was performed by a physician. Thyroid volume was determined by ultrasound, using the ellipsoid method (20). All examiners were blinded to the ThyPRO results. The patients were classified according to primary diagnosis, i.e. their initial diagnosis, prior to treatment. For example, patients with a non-toxic goitre, who had their thyroid removed and thus have hypothyroidism and receive thyroid hormone replacement, were classified with a diagnosis of non-toxic goitre. Self-completed data were entered using optical scanning. All clinical data were entered via SPSS Data Entry Builder 4.0 (SPSS Inc, Chicago, IL, USA) by medical staff. Data were converted into

SAS datasets, and all analyses were performed with SAS 9.1 (SAS Institute, Cary, NC, USA). The project was approved by the local ethical committee (KF01 2006-1579) and the Danish data protection agency and registered at ClinicalTrials.gov (NCT00150033).

Outcome

The ThyPRO questionnaire is self-administered and measures QoL with 13 scales (see Table 1), covering physical and mental symptoms, well-being and function as well as impact of thyroid disease on participation (i.e. social and daily life) and overall QoL. It consists of 84 items and, on average, takes 14 min to complete. Each scale ranges 0–100 with increasing scores indicating decreasing QoL (i.e. more symptoms or greater impact of disease) (16).

Known groups

Groups with expected high and low scores were defined *a priori* by a panel of four thyroid experts. The criteria used for this classification is outlined in Table 1. They include clinical data, physician ratings of overall clinical condition, and depression and anxiety status according to Hospital Anxiety and Depression Scale (HADS). The overall physician ratings were obtained at the out-patient visit, where the consulting physician, who was

blinded to the ThyPRO results, rated the overall clinical condition of the patient on a five-point scale ranging from 'very bad' to 'excellent', based on all available clinical information, including patient history. Depression or anxiety according to HADS was scored using its standard thresholds: a score <8 on the depression scale indicates absence of depression, and a score above 10 indicates depression, and likewise for the anxiety scale. In addition, we evaluated the correlation between the ThyPRO depressivity and anxiety scales and the HADS depression and anxiety scales respectively using both parametric (Pearson) and non-parametric (Spearman) correlation.

Test–retest data

Since fewer patients were required for test–retest analyses than for the other validation analyses, only patients enrolled at RH during a limited period (mid-April to mid-June 2007) were included in the test–retest study. The only difference compared to the main study was the fact that they were asked to complete another questionnaire about 2 weeks after their first response. In addition, they rated any change occurring since the initial response on a seven-point scale ('compared to when you answered the first time, would you say your overall state is much worse/somewhat worse/a little worse/more or less the same/a little better/somewhat

Table 1 Description of the expected high versus low score groups used in the known groups comparisons.

Scale	Expected high score group	Expected low score group
Goitre symptoms	Patients with untreated non-toxic diffuse or multinodular goitre (<i>n</i> =105)	Patients with non-goitrous autoimmune hypothyroidism treated with L-thyroxine for at least 3 months (<i>n</i> =107)
Hyperthyroid symptoms	Patients with hyperthyroid Graves' disease or nodular goitre and overt hyperthyroidism (<i>n</i> =70)	Untreated non-toxic diffuse or nodular goitre (<i>n</i> =161)
Hypothyroid symptoms	Patients diagnosed with overt hypothyroidism within the last 3 months (<i>n</i> =20)	Untreated non-toxic diffuse or nodular goitre (<i>n</i> =161)
Eye symptoms	Patients with TAO and NOSPECS > 1 (i.e. worse than 'Only signs') (<i>n</i> =16)	Untreated non-toxic diffuse or nodular goitre (<i>n</i> =161)
Tiredness	Patients with overall clinical condition rated by physician as 'very bad' (<i>n</i> =12)	Patients with overall clinical condition rated by physician as 'excellent' (<i>n</i> =88)
Cognitive impairment	Patients with overall clinical condition rated by physician as 'very bad' (<i>n</i> =12)	Patients with overall clinical condition rated by physician as 'excellent' (<i>n</i> =87)
Anxiety	Patients with a HADS anxiety score indicating anxiety (i.e. score > 10) (<i>n</i> =146)	Patients with a HADS anxiety score indicating no anxiety (i.e. score < 8) (<i>n</i> =577)
Depressivity	Patients with a HADS depression score indicating depression (i.e. score > 10) (<i>n</i> =64)	Patients with a HADS depression score indicating no depression (i.e. score < 8) (<i>n</i> =703)
Emotional susceptibility	Patients with overall clinical condition rated by physician as 'very bad' (<i>n</i> =12)	Patients with overall clinical condition rated by physician as 'excellent' (<i>n</i> =87)
Impaired social life	Patients with overall clinical condition rated by physician as 'very bad' (<i>n</i> =12)	Patients with overall clinical condition rated by physician as 'excellent' (<i>n</i> =84)
Impaired daily life	Patients with overall clinical condition rated by physician as 'very bad' (<i>n</i> =12)	Patients with overall clinical condition rated by physicians as 'excellent' (<i>n</i> =85)
Impaired sex life	Patients with overall clinical condition rated by physician as 'very bad' (<i>n</i> =9)	Patients with overall clinical condition rated by physician as 'excellent' (<i>n</i> =85)
Cosmetic complaints	Patients with TAO and NOSPECS above 1 and patients with non-toxic diffuse or multinodular goitre with a volume greater than 150 ml (<i>n</i> =21)	Patients with autoimmune hypothyroidism and an overall clinical condition rated by physician as 'excellent' (<i>n</i> =22)

TAO, thyroid-associated ophthalmopathy; NOSPECS, a classification system for defining severity of TAO; HADS, Hospital Anxiety and Depression Scale.

better/much better'). Patients who completed the second questionnaire between 10 and 24 days after the first, and who rated themselves as stable were included in test–retest analyses.

Statistical analyses

Differences in mean scale scores between the expected high versus low level groups were analysed with Student's unpaired *t*-test, with Satterthwaites correction in case of unequal variances according to the folded *F* test, using SAS PROC TTEST.

Test–retest reliability was evaluated by intra-class correlations between the two measurements (21–24). Correlations were calculated using SAS PROC GLM, and 95% confidence intervals were estimated by empirical bootstrap (25, SV Thorsen and JB Bjorner, unpublished observations).

Results

In total, 907 responses from 1316 eligible patients were obtained, yielding an overall response rate of 69%, as detailed elsewhere (16). Of these, 195 were included in the test–retest study, 149 (76%) of whom returned a second response. Eighty-seven of the 195 fulfilled the criteria for inclusion in the analyses (Fig. 1).

Clinical characteristics for the total sample as well as for the test–retest subsample are given in Table 2.

As seen in Fig. 2, the expected high level groups had substantially higher mean scores compared with the expected low level groups on all ThyPRO scales.

Figure 1 Derivation of the test–retest subsample.

Table 2 Patient characterization. Clinical characteristics are presented for the total sample and for the test–retest subsample.

	Total sample	Test–retest subsample
Sample size	907	87
Women (%)/men	787 (87)/120	81 (93)/6
Age (mean (s.d.))	51 (15)	49 (13)
Diagnosis (n (%))		
Diffuse non-toxic goitre	18 (2)	1 (1)
Multinodular non-toxic goitre	154 (17)	12 (14)
Uninodular non-toxic goitre	68 (7)	2 (2)
Solitary cyst	19 (2)	
Multinodular toxic goitre	108 (12)	3 (3)
Uninodular toxic goitre	37 (4)	5 (6)
Graves' disease	168 (19)	28 (32)
TAO	94 (10)	9 (10)
Autoimmune hypothyroidism	199 (22)	24 (28)
Subacute thyroiditis	9 (1)	
Post-partum thyroiditis	8 (1)	1 (1)
Other thyroid disease	25 (3)	2 (2)
Months since diagnosis (median (range)) ^a	27 (–0.9–607)	47 (0.7–269)
Thyroid treatment (n (%))		
No thyroid treatment (ever)	283 (31)	10 (11)
Antithyroid medication	162 (18)	20 (23)
L-thyroxine	292 (32)	43 (50)
Radioiodine	114 (13)	14 (16)
Thyroidectomy	132 (14)	8 (10)
Other treatment	4 (0.4)	0
Current thyroid function ^b (n (%))		
Euthyroid	530 (58)	60 (70)
Mildly hypothyroid	124 (14)	8 (9)
Overtly hypothyroid	15 (2)	0
Mildly hyperthyroid	122 (13)	16 (19)
Overtly hyperthyroid	98 (11)	2 (2)
Positive TPOAb (n (%))	367 (46)	55 (64)
Positive TSHRAb (n (%))	174 (35)	22 (37)
Thyroid volume (mean (s.d.), ml)	26 (32)	25 (25)

^aNegative durations reflect patient responding to the questionnaire before a final thyroid diagnosis was established.

^bEuthyroid, normal TSH; mildly hypothyroid, elevated TSH and normal T₄; overtly hypothyroid, elevated TSH and decreased T₄; mildly hyperthyroid, decreased TSH and normal T₄ and T₃; overtly hyperthyroid, decreased TSH and elevated T₄ or T₃.

All differences were statistically significant using unpaired *t*-tests ($P < 0.001$ for all scales except hypothyroid symptoms and cosmetic complaints, where $P < 0.05$). The correlations between ThyPRO and HADS scales were 0.72 for depression/depression and 0.77 for anxiety ($P < 0.0001$ for both), both with parametric and non-parametric methods.

In the test–retest analyses, all intra-class correlations were above 0.70 (Table 3) and all but two (anxiety (0.77) and hypothyroid symptoms (0.80)) scales were above the 'almost perfect' concordance (i.e. > 0.81) (26).

Discussion

The purpose of the present study was to investigate clinical known-groups validity and test–retest reliability of the Danish version of the thyroid-specific QoL questionnaire, the ThyPRO.

Figure 2 Mean scale scores for the expected high score groups versus the low score groups for the 13 ThyPRO scales. Vertical lines indicate 95% confidence intervals of the means.

When comparing specified subgroups of patients expected to have high scale scores on each of the 13 QoL scales to subgroups expected to have low scores, we found significant differences on all scales, supporting the clinical validity of the ThyPRO. This is an important and encouraging finding.

The magnitudes of these differences varied: differences on the symptom scales, the cosmetic complaints scale and the impaired social life scale were smaller than the rest. The simple conclusion that these scales are just less sensitive (i.e. are less efficient in detecting differences among groups) can be elaborated by a closer look at the criteria used to define these groups. In fact, the analyses of these scales differed from the rest, in that the criteria defining the high versus low score groups were entirely clinical and biochemical, except for the impaired daily life scale. Criteria for the other scales were based on the consulting clinician's rating of the overall clinical condition or HADS anxiety and depression scores. To better understand these differences, it is useful to consider a theoretical model for health outcomes. Based on the WHO ICF framework (27) and on the conceptual model for health outcomes proposed by Wilson & Cleary (28), we propose to distinguish four levels of health outcomes: i) biological and physiological variables, ii) symptoms and signs, iii) functioning, i.e. what you can physically do, and iv) participation, i.e. how you function in your social environment. Regarding the symptom scales, the criteria used to define the groups for comparisons relate to biological and physiological variables, whereas the scale scores reflect patients' experiences of symptoms and signs. The same is true for the cosmetic complaints scale. Thus, since the disease group comparisons do not directly concern the same

level of health outcomes as the ThyPRO scales, it is reasonable that the associations are fairly small. On the other hand, the overall clinical rating used for evaluating the other scales, except anxiety and depression, is likely to be based on symptoms and signs, and functioning, as well as diagnosis and physiological measures. Thus, strong associations for these scales are likely, which is also what we found, except for the impaired social life scale, which reflects participation rather than functioning. Finally, we are likely to see strong associations between ThyPRO scales for anxiety and depression, and other scales assessing the same domains, i.e. HADS.

For all scales, test-retest reliability was above standard thresholds for adequate reliability, indicating that the measures have appropriate reliability for use in clinical studies (29). Since the test-retest reliabilities found in this study were a little lower than the

Table 3 Intra-class correlation coefficients with 95% confidence intervals (CI) between the two measurements in the test-retest subsample.

Scale	Intra-class correlation coefficient (95% CI)
Goitre symptoms	0.87 (0.81–0.91)
Hyperthyroid symptoms	0.89 (0.82–0.93)
Hypothyroid symptoms	0.80 (0.71–0.87)
Eye symptoms	0.86 (0.77–0.92)
Tiredness	0.85 (0.77–0.91)
Cognitive complaints	0.88 (0.79–0.93)
Anxiety	0.77 (0.64–0.88)
Depressivity	0.88 (0.82–0.93)
Emotional susceptibility	0.87 (0.80–0.91)
Impaired social life	0.84 (0.72–0.91)
Impaired daily life	0.83 (0.71–0.91)
Impaired sex life	0.86 (0.74–0.94)
Cosmetic complaints	0.85 (0.79–0.91)

coefficients found by internal consistency reliability (16), the test–retest reliabilities may even underestimate true reliability slightly, possibly due to less than perfect clinically stable conditions of the patients in the test interval. In fact, the mean score was indeed a little lower at time of re-evaluation, for all scales. No scale achieved a reliability coefficient above the traditional cut-point of 0.90 required for scales used with individual patients.

Since this is the first QoL measure for use in thyroid patients covering the majority of benign thyroid disorders, no previous studies evaluating measurement properties of such measures exist. Regarding individual diagnoses, test–retest reliability and clinical known groups validity have been evaluated (6) for a disease-specific questionnaire for patients with TAO. Test–retest reliability estimates were comparable to the ones found in the present study. However, their clinical validity analyses did not entirely support the validity of the questionnaire, since some of the expected associations with clinical variables were not found (6). The interpretations of these lacking associations offered by the authors are analogous to our interpretations regarding the criteria for evaluations of the symptom scales. They suggested that it was due to the fact that the clinical variables defining the groups related to health outcomes (i.e. biological and physiological variables) differed from the health outcomes relating to QoL (i.e. symptoms, functioning and participation). For another TAO questionnaire, no associations with clinical variables were found at all (8). To our knowledge, no other study has evaluated clinical validity or test–retest reliability of any of the other existing thyroid-specific (i.e. for individual diagnoses) questionnaires (reviewed in (13)).

Among the strengths of our study is a large sample size, enabling a detailed analysis of each group of benign thyroid disorders with adequate power. Moreover, the patients are clinically well characterized, especially compared to other studies of generic or specific QoL in thyroid patients (13). The fact that each patient was rated by a physician at about the same time that they completed the questionnaire is an important strength of the clinical validity analyses. This is also true for the availability of a ‘stability-measure’ for identification of patients to enter the test–retest substudy. Another strength is the fact that we used only few exclusion criteria, ensuring validity of the questionnaire also for subpopulations with e.g. psychiatric co-morbidities, etc. Of course, future descriptive studies evaluating HRQL in thyroid patients may consider excluding these patient groups in order to limit the description to only impairments related to thyroid diseases.

One limitation of the study relates to its design: due to our cross-sectional design, our analyses of clinical validity are limited to cover differences between groups. Ongoing studies are evaluating the ability of the questionnaire to detect relevant changes in patients over time (responsiveness) and the minimal important differences in scores. However, compared to other

disease-specific instruments, including those covering other disease areas than thyroid diseases, the ThyPRO is already very well validated. Another limitation relates to the clinician evaluation. Although it is of value that these are available for each individual patient at time of survey completion, the fact that they are also based on patient history implies that they are not entirely external criteria, as are for example thyroid function tests, etc. Still, they do represent evaluations external to the ThyPRO, and clinician evaluations are often used in studies of criterion validity (30, 31).

Although the patients are clinically well characterized, and a large number of clinical descriptors are available at time of completion, the timing is not perfect; some patients may have had blood samples drawn up to about 3 weeks away from their questionnaire response. However, this would weaken the associations between e.g. thyroid function tests and HRQL data, but we still find positive associations despite such hypothetical bias.

Information from clinical known-groups studies, responsiveness studies and item response theory analyses could be utilized in an attempt to develop shorter versions of the ThyPRO by identifying the best performing items to be included in such abbreviated outcome measures. Ongoing studies are ensuring cross-cultural validity of this measure in seven languages, including English, and versions in further languages are being planned.

We recommend the use of the ThyPRO measure in studies evaluating important clinical questions regarding therapy of thyroid patients, such as whether patients with mild thyroid disease benefit from treatment (32) and whether block replacement therapy with antithyroid drugs and levothyroxine is associated with a better QoL in patients with hyperthyroidism than monotherapy with antithyroid drugs (33).

In conclusion, all scales of the ThyPRO detected clinically relevant differences among thyroid patients, and the ThyPRO was found to have adequate test–retest reliability.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding

This study has been supported by grants from the Danish Medical Research Council, Agnes and Knut Mørk's Foundation, Aase and Ejnar Danielsen's Foundation, Else and Mogens Wedell-Wedellsborg's Foundation, the Genzyme Corporation, the Novo Nordisk Foundation and the Danish Thyroid Foundation.

Acknowledgements

We wish to express our gratitude to the staffs and colleagues at Department of Endocrinology at Rigshospitalet as well as at Odense University Hospital.

References

- Guyatt GH, Feeny DH & Patrick DL. Measuring health-related quality of life. *Annals of Internal Medicine* 1993 **118** 622–629.
- Sloan JA, Frost MH, Berzon R, Dueck A, Guyatt G, Moinpour C, Sprangers M, Ferrans C & Cella D. The clinical significance of quality of life assessments in oncology: a summary for clinicians. *Supportive Care in Cancer* 2006 **14** 988–998.
- Velikova G, Booth L, Smith AB, Brown PM, Lynch P, Brown JM & Selby PJ. Measuring quality of life in routine oncology practice improves communication and patient well-being: a randomized controlled trial. *Journal of Clinical Oncology* 2004 **22** 714–724.
- Fahrenfort JJ, Wilterdink AM & van der Veen EA. Long-term residual complaints and psychosocial sequelae after remission of hyperthyroidism. *Psychoneuroendocrinology* 2000 **25** 201–211.
- Terwee CB, Gerding MN, Dekker FW, Prummel MF & Wiersinga WM. Development of a disease specific quality of life questionnaire for patients with Graves' ophthalmopathy: the GO-QOL. *British Journal of Ophthalmology* 1998 **82** 773–779.
- Terwee CB, Gerding MN, Dekker FW, Prummel MF, van der Pol JP & Wiersinga WM. Test–retest reliability of the GO-QOL: a disease-specific quality of life questionnaire for patients with Graves' ophthalmopathy. *Journal of Clinical Epidemiology* 1999 **52** 875–884.
- Terwee CB, Dekker FW, Mourits MP, Gerding MN, Baldeschi L, Kalmann R, Prummel MF & Wiersinga WM. Interpretation and validity of changes in scores on the Graves' ophthalmopathy quality of life questionnaire (GO-QOL) after different treatments. *Clinical Endocrinology* 2001 **54** 391–398.
- Tehrani M, Kruppenauer F, Mann WJ, Pitz S, Dick HB & Kahaly GJ. Disease-specific assessment of quality of life after decompression surgery for Graves' ophthalmopathy. *European Journal of Ophthalmology* 2004 **14** 193–199.
- Jaeschke R, Guyatt G, Gerstein H, Patterson C, Molloy W, Cook D, Harper S, Griffith L & Carbotte R. Does treatment with L-thyroxine influence health status in middle-aged and older adults with subclinical hypothyroidism? *Journal of General Internal Medicine* 1996 **11** 744–749.
- Jaeschke R, Guyatt G, Cook D, Harper S & Gerstein HC. Spectrum of quality of life impairment in hypothyroidism. *Quality of Life Research* 1994 **3** 323–327.
- Saravanan P, Chau WF, Roberts N, Vedhara K, Greenwood R & Dayan CM. Psychological well-being in patients on 'adequate' doses of L-thyroxine: results of a large, controlled community-based questionnaire study. *Clinical Endocrinology* 2002 **57** 577–585.
- McMillan CV, Bradley C, Woodcock A, Razvi S & Weaver JU. Design of new questionnaires to measure quality of life and treatment satisfaction in hypothyroidism. *Thyroid* 2004 **14** 916–925.
- Watt T, Groenvold M, Rasmussen AK, Bonnema SJ, Hegedüs L, Bjorner JB & Feldt-Rasmussen U. Quality of life in patients with benign thyroid disorders. A review. *European Journal of Endocrinology* 2006 **154** 501–510.
- Watt T, Hegedüs L, Rasmussen AK, Groenvold M, Bonnema SJ, Bjorner JB & Feldt-Rasmussen U. Which domains of thyroid-related quality of life are most relevant? Patients and clinicians provide complementary perspectives *Thyroid* 2007 **17** 647–654.
- Watt T, Rasmussen AK, Groenvold M, Bjorner JB, Watt SH, Bonnema SJ, Hegedüs L & Feldt-Rasmussen U. Improving a newly developed patient-reported outcome for thyroid patients, using cognitive interviewing. *Quality of Life Research* 2008 **17** 1009–1017.
- Watt T, Bjorner JB, Groenvold M, Rasmussen AK, Bonnema SJ, Hegedüs L & Feldt-Rasmussen U. Establishing construct validity for the thyroid-specific patient reported outcome measure (ThyPRO): an initial examination. *Quality of Life Research* 2009 **18** 483–496.
- Streiner DL & Norman GR. Health measurement scales. In *A Practical Guide to their Development and Use*, edn 4. Oxford: Oxford University Press, 2008.
- Werner SC. Modification of the classification of the eye changes of Graves' disease: recommendations of the Ad Hoc Committee of the American Thyroid Association. *Journal of Clinical Endocrinology and Metabolism* 1977 **44** 203–204.
- Mourits MP, Prummel MF, Wiersinga WM & Koornneef L. Clinical activity score as a guide in the management of patients with Graves' ophthalmopathy. *Clinical Endocrinology* 1997 **47** 9–14.
- Hegedüs L. Thyroid ultrasound. *Endocrinology and Metabolism Clinics of North America* 2001 **30** 339–360.
- Tammemagi MC, Frank JW, Leblanc M, Artsob H & Streiner DL. Methodological issues in assessing reproducibility – a comparative study of various indices of reproducibility applied to repeat ELISA serologic tests for Lyme disease. *Journal of Clinical Epidemiology* 1995 **48** 1123–1132.
- Shrout PE & Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 1979 **86** 420–428.
- Deyo RA, Diehr P & Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Controlled Clinical Trials* 1991 **12** 142S–158S.
- Fayers PM & Machin D. *Quality of Life: Assessment, Analysis and Interpretation* Chichester: John Wiley and Sons, 2000.
- Henderson AR. The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. *Clinica Chimica Acta* 2005 **359** 1–26.
- Landis JR & Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977 **33** 159–174.
- World Health Organization. *International Classification of Functioning, Disability, and Health* Geneva: World Health Organization, 2001.
- Wilson IB & Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *Journal of the American Medical Association* 1995 **273** 59–65.
- Nunnally JC & Bernstein I. *Psychometric Theory*, edn 3. New York: McGraw-Hill, 1994.
- Schatz M, Sorkness CA, Li JT, Marcus P, Murray JJ, Nathan RA, Kosinski M, Pendergraft TB & Jhingran P. Asthma control test: reliability, validity, and responsiveness in patients not previously followed by asthma specialists. *Journal of Allergy and Clinical Immunology* 2006 **117** 549–556.
- Miller IW, Epstein NB, Bishop DS & Keitner GI. The McMaster family assessment device: reliability and validity. *Journal of Marital and Family Therapy* 2009 **11** 345–356.
- Biondi B & Cooper DS. The clinical significance of subclinical thyroid dysfunction. *Endocrine Reviews* 2008 **29** 76–131.
- Hegedüs L. Treatment of Graves' hyperthyroidism: evidence-based and emerging modalities. *Endocrinology and Metabolism Clinics of North America* 2009 **38** 355–371.

Received 25 September 2009

Accepted 1 October 2009