Psychology Health & Medicine – Methodological issue

Title page

Title: **Establishing the usefulness of the GO-QOL in a UK hospital-treated population with thyroid eye disease in the CIRTED trial**

Corresponding author details:
Dr Sue Jackson, Centre for Appearance Research,
Department of Psychology, University of the West of England, Frenchay Campus,
Coldharbour Lane, Frenchay, Bristol BS16 1QY
Tel: 01454 250482 Email: hellosue@suejackson.me.uk
ORCID: 0000-0002-1684-2224

Author details:
Alina Dietrich, Cardiff University School of Medicine,
UHW Main Building, Heath Park, Cardiff, CF14 4XN, UK
Tel: +44 (0) 75813 29499 Email: DietrichAM@cardiff.ac.uk

Peter Taylor, Thyroid Research Group, Systems Immunity Research Institute, Cardiff University School of Medicine, Cardiff, UK
C2 Link corridor, Cardiff University, Heath Park, Cardiff CF14 4XN
Tel: 07590 520741 Email: taylorpn@cardiff.ac.uk
ORCID: 0000-0002-3436-422X

Paul White, University of the West of England
Associate Professor (Applied Statistics), Director Applied Statistics Group, University of the West of England, Frenchay Campus, Coldharbour Lane, Frenchay, Bristol BS16 1QY
Tel: +44 117 32 83777 Email: Paul.White@uwe.ac.uk
ORCID: 0000-0002-7503-9896

Victoria Wilson, Bristol Eye Hospital, University of Bristol
Faculty of Health Sciences, University of Bristol, Tyndall Avenue, Bristol BS8 1TH, UK
Tel: 0117 331 4555 Email: victoria.wilson@bristol.ac.uk

Jimmy Uddin, Moorfields Eye Hospital,
Moorfields Eye Hospital, City Road, London, EC1V 2PD, UK
Tel: 020 7253 3411 Email: jimmy.uddin@moorfields.nhs.uk

Richard William John Lee, University of Bristol and NIHR Moorfields Biomedical Research Centre
Faculty of Health Sciences, University of Bristol, Tyndall Avenue, Bristol BS8 1TH, UK
Tel: +44 (0)117 331 1479 Email: richard.lee@bristol.ac.uk

Colin Dayan, Cardiff University School of Medicine
Professor of Clinical Diabetes and Metabolism**,** C2 Link corridor, Cardiff University, Heath Park, Cardiff CF14 4XN
Tel: +44 (0)29 2074 2182 Email: Email: DayanCM@cardiff.ac.uk

Dr Sue Jackson, Centre for Appearance Research,
Department of Psychology, University of the West of England, Frenchay Campus,
Coldharbour Lane, Frenchay, Bristol BS16 1QY
Tel: 01454 250482 Email: hellosue@suejackson.me.uk
ORCID: 0000-0002-1684-2224

On behalf of the CIRTED investigators

Word Count: 3,158
Word Limit: 4,000 (excluding abstract, tables, figures & references)

Abstract

Thyroid eye disease (TED) is a potentially sight-threatening and cosmetically disfiguring condition arising in 25-50% of patients with Graves' hyperthyroidism. CIRTED is the first study to evaluate the long-term role of radiotherapy and prolonged immunosuppression with azathioprine in treating TED, one aim of which was to validate the use of the English version of GO-QOL in an UK population with TED. In a three stage design over a 48 week period, the GO-QOL was tested and compared to a general measure of quality of life (WHOQOL-Bref). In stage 1 utilising a standard 14 day test-retest design both GO-QOL subscales achieved Cronbach's alphas demonstrating excellent validity and internal reliability (Visual Function 0.929 and 0.931; Appearance 0.888 and 0.906). In stage 2, Repeated Measures ANOVA demonstrated longitudinal validity, with both subscales of the GO-QOL showing significant change over time (Visual Function, η^2=0.114, p<.001; Appearance, η^2=0.069, p<.002). In stage 3 the GO-QOL showed discriminant validity at the week 48 time point, with the visual function subscale being able to detect changes in groups identified by clinicians (using BCCOM ratings of improvement or deterioration), while both subscales could detect group differences when based on participants' subjective ratings of TED noticeability and severity. The results of this project provide support for the English translation of the GO-QOL as an outcome measure for patients with moderately severe active Graves' orbitopathy/TED.

Word Count: 224
Word Limit: 300 words

Keywords: thyroid eye disease (TED), Graves' ophthalmopathy (GO), GO-QOL, adults, psychometrics, reliability, validity

Introduction

Graves' ophthalmopathy or orbitopathy (GO) also known as thyroid eye disease (TED) is a potentially sight-threatening and cosmetically disfiguring condition arising in 25-50% of patients with Graves' hyperthyroidism (Bahn & Heufelder, 1993). The condition is rare and causes redness and grittiness of the eye and can lead to disfiguring swelling of the eyelids, proptosis (abnormal protrusion or displacement of the eye) and even blindness (Weetman, 1991). GO is an autoimmune disorder, linked to thyroid

autoimmunity by autoantigens shared between the thyroid and the orbit of the eye (Perros, Crombie & Kendall-Taylor, 1995). Inflammatory processes are activated and fibroblasts in the eyes' orbital tissue become stimulated leading to orbital tissue swelling, hyaluronan production, and expansion of the extraocular muscles, retro-orbital fat and connective tissues (Khoo & Bahn, 2007).

Currently the management of GO/TED is considered suboptimal, and available treatments do not specifically target the underlying pathogenic process (Bartalena et al, 2016). The Combined Immunosuppression and Radiotherapy in Thyroid Eye Disease (CIRTED) trial was designed to assess the effect of using radiotherapy and the immunosuppressive drug azathioprine in combination with standard prednisolone treatment (Rajendram et al, 2008). CIRTED is the first study to evaluate the long-term role of radiotherapy and prolonged immunosuppression with azathioprine in treating GO/TED.

It is well established that TED can have a major impact on quality of life, in particular disfiguring changes to the eyes and face which can have a direct impact on psychological health (Coulter, Frewin, Krassas & Perros, 2007). As the aim of treating GO/TED is to improve patients' visual function as well as making them look and feel better, it is important to assess the patients' perception of these markers as part of a clinical trial. The GO-QOL questionnaire was developed by Terwee, Gerding, Dekker, Prummel & Wiersinga (1998) as a TED specific quality of life questionnaire that can be used as an outcome measure for studies and may also be of use in clinical practice. Marcocci et al (2011) tested the use of selenium in mild GO and used the GO-QOL to evaluate quality of life outcomes. They showed a correlation that indicated that as participants' improved with selenium treatment their quality of life also improved, as measured by the GO-QOL.

The aim of this three stage project was to validate the use of the English version of GO-QOL in an UK population with TED. The GO-QOL is available in both English and Dutch; the Dutch version having previously been validated in the Netherlands (Terwee et al, 1999). In the first stage of the work, the internal validity and test-retest study, the objective was to assess the consistency of the GO-QOL in measuring functional and

appearance-related issues resulting from TED over a 14 day period where the expectation is that scores on both administrations should be correlated.

In the second stage, we measured longitudinal validity i.e. the responsiveness of the GO-QOL to changes in TED post-treatment against a more general measure of quality of life (in this case, the WHOQOL-Bref; The WHOQOL Group, 1998). This work had two aspects:

1. If the GO-QOL is valid (i.e. sensitive to changes in visual functioning and appearance as a result of TED) we would expect larger effect sizes for the GO-QOL than for the general quality of life measure (WHOQOL-Bref).

2. Furthermore, changes in clinical characteristics relating to visual functioning and appearance, as indicated by transitional variables, should be more closely associated with changes in scores on the relevant subscales of the GO-QOL than with the WHOQOL-Bref.

The third and final stage was to explore the extent to which the GO-QOL can demonstrate discriminant validity. That is, whether the GO-QOL can distinguish between patient populations based on either clinician ratings of improvement or participants' subjective measures. Although it should be noted that subjective severity does not always correlate well with objective measures of disease and physicians' assessments (Bessell, Dures, Semple & Jackson, 2012).

Methods

Design

The study protocol has been described in detail previously (Rajendram et al, 2008). In brief, CIRTED is a 2x2 factorial design, double-masked, multi-centre, randomized controlled trial (see Figure 1 for a diagram of the trial design).

Ethical approval

Ethical approval was from the UK's National Health Service South West Central Bristol Research Ethics Committee (REC reference: 05/Q2006/62). Clinical Trial Authorisation was given by the Medicines and Healthcare products Regulatory Agency (MHRA, reference:

03299/0003/001-0001; ISRCTN22471573) with the University of Bristol acting as the legal sponsor. Research governance and local Research and Development approvals were obtained across all sites prior to the start of recruitment. All participants gave written informed consent.

[Figure 1 near here]

Materials (see also Table 1)

The Graves' ophthalmopathy quality of life assessment (GO-QOL) questionnaire consists of two subscales, each comprising eight questions, on visual function and the psychological impact of changed appearance (Terwee et al, 1998).

The WHOQOL-Bref is a widely used and previously validated measure of quality of life consisting of 28 items which cover subjective overall quality of life and subjective overall health, plus items relating to domains of physical, psychological, social relationships and environment (The WHOQOL Group, 1998). The WHOQOL-Bref has been widely used with a range of populations and is reported to have good psychometric properties (Skevington, Lotfy & O'Connell, 2004).

Two transitional variables relating closely to the subscales of the GO-QOL were included at follow up as an external standard to identify changes post-treatment (labelled "T1" and "T2"). These variables were agreed with the authors of the GO-QOL as being suitable for this purpose. In order to explore the impact of TED and its treatment on psychological adjustment and daily functioning more broadly, Visual Analogue Scales (VAS) were also included in the study. VAS scales are easy for respondents to complete and are often used in clinical assessments (Carr, 1997).

Clinician ratings of disease severity and activity were also included. The Binary Composite Clinical Outcome Measure (BCCOM) a system of major and minor criteria used in previous TED trials (Prummel et al, 2004; Marcocci et al, 2001; Mourits et al, 2000). It is a clinician rating of improvement in the CIRTED trial used at 1 year post treatment to classify study participants' treatment as being successful or not (see Appendix 1 for information on its derivation).

[Table 1 near here]

Protocol & Participants

All participants referred to a trial centre during the duration of the study were considered for inclusion. Participants were prescribed a high dose of tapering prednisolone at their initial enrolment visit. If they were eligible, responded to steroids and not excluded, participants were randomized into one of four trial arms (see Figure 1, Table 2, and CIRTED study protocol – Rajendram et al, 2008).

For validating test-retest reliability, participants completed the GO-QOL twice at a two-week interval, time point 1 (-2 weeks, i.e. enrolment into the study) and time point 2 (0 weeks, i.e. randomization into the study). Two weeks have previously been used by Terwee et al (1999) for assessing test-retest reliability, as it is long enough to avoid recall bias and short enough for patients not to experience clinically important changes in their condition. Data from participants for the longitudinal validity testing was collected at all four time points.

[Table 2 near here]

Results

The results are reported in the three stages that the work was undertaken. All analyses were undertaken using SPSS version 23.

Stage 1 GO-QOL validation: internal validity and test re-test

167 participants attended for study enrolment, while 133 attended the randomization appointment. Of these, 142 participants completed the GO-QOL at study enrolment, and 126 completed the GO-QOL at randomization (see Table 3).

[Table 3 near here]

Item response tables for time points 1 (-2 weeks, enrolment) and 2 (0 weeks, randomisation) were generated to identify patterns of responses and are shown in Tables 4 & 5.

[Tables 4 & 5 near here]

Not all participants cycled or drove, hence the lower number of responses to questions 1 and 2 on the Visual Function subscale at both time points. As per the questionnaire authors' instructions (Terwee et al, 1998), adjustments for missing data were made when totalling the raw scores for each subscale prior to transforming them into a total score out of 100 for subsequent analyses.

Internal validity was assessed using Cronbach's alpha calculations; at time 1 (-2 weeks, enrolment) the visual function subscale achieved a Cronbach's alpha score of 0.929 (CI 0.909 – 0.944) while the appearance subscale recorded 0.888 (CI 0.857 – 0.912). At time 2 (week 0, randomization), the alpha results for visual function remained virtually unchanged at 0.931 (CI 0.912 – 0.946), while the appearance subscale improved slightly to 0.906 (CI 0.880 – 0.926). This indicates good internal validity of the subscales at both time points, with values over 0.7 generally acceptable for psychometric questionnaires (BPS, 1992).

Pearson correlation coefficients for both subscales were found to be highly significant (visual function, $r=0.774$, $p<.001$; appearance, $r=0.862$, $p<.001$), indicating the robust test-retest reliability of the GO-QOL subscales.

Stage 2: Longitudinal validation of GO-QOL

Longitudinal validation was performed using data from the 126 participants that were randomised into the trial (Table 6).

[Table 6 near here]

In this second stage, we measured the responsiveness of the GO-QOL to changes in TED post-treatment against the WHOQOL-Bref (a more general measure of quality of life). We hypothesized that, if the GO-QOL is sensitive to changes in visual functioning and appearance as a result of TED we would expect larger effect sizes for the GO-QOL than for the general measure (WHOQOL-Bref). Effect size was quantified using partial eta squared where the following thresholds can be used to determine effect size interpretation: η^2 of less than 0.01 indicates an inconsequential effect, η^2 between 0.01

and 0.09 a small effect, η^2 between 0.09 to 0.25 a medium effect, η^2 from 0.25 to 0.50 a large effect, and η^2 over 0.50 a very large effect (Cohen, 1988).

Of the 126 randomized participants, 108 participants provided enough completed questionnaire data at the 12 week trial appointment, while 100 of those participants who attended the 48 week appointment completed the study measures (see Table 7).

[Table 7 near here]

For the sake of completeness, internal validity was assessed again using Cronbach's alpha calculations; at both time 3 (short term 12 week trial appointment) and time 4 (long term 48 week appointment). At time 3 the visual function subscale achieved a Cronbach's alpha score of 0.904 (CI 0.854 – 0.937) while the appearance subscale recorded 0.918 (CI 0.891 – 0.938). At time 4 the alpha result for visual function had reduced slightly to 0.887 (CI 0.823 – 0.928), while the appearance subscale remained largely unchanged at 0.915 (CI 0.886 – 0.937). As before, with all values over 0.7 this indicates good internal validity of the subscales at both time points.

Repeated measures ANOVA (with time as 3 level factor, i.e. -2, 12, and 48 weeks) revealed significant changes over time for both GO-QOL subscales (Visual Function, p=0.001, Appearance, p=0.002), and for the psychological domain of the WHOQOL (p=0.002). Effect sizes varied between the subscales for the GO-QOL with only Visual Function recording a medium effect size (η^2=0.114); the Appearance subscale showed a small effect size (η^2=0.069). Effect sizes for the WHOQOL subscales were all small (psychological, η^2=0.064; physical, η^2=0.037; social, η^2=0.041; environment, η^2=0.043) (Table 8).

[Table 8 near here]

We also hypothesized that changes in clinical characteristics relating to visual functioning and appearance, as indicated by the transitional variables, should be more closely associated with changes in scores on the relevant subscales of the GO-QOL than with the WHOQOL-Bref. As expected since all study participants were taking steroid treatment, correlation calculations showed positive correlations between both the

transitional variables (T1 and T2) and the GO-QOL appearance subscale at the 12-week time point, while the Visual Function subscale only correlated with T1 (Table 9). Similarly, three of the WHOQOL-Bref subscales significantly correlated with T1, while all four subscales significantly correlated with T2. At 48-weeks significant correlations were observed only for the appearance subscale of the GO-QOL with the transitional variables.

We included two Visual Analogue Scales to explore the impact of treatment for TED on psychological adjustment and daily functioning more broadly. At 12 weeks significant correlations were observed between both VAS scales and the appearance subscale of the GO-QOL. Three of the four domains of the WHOQOL-Bref significantly correlated with the VAS for participant perceived noticeability of TED (psychological, physical and environment), while a different trio of domains significantly correlated with the VAS for participant ratings of severity of TED (physical, social and environment). At 48 weeks, the only significant correlations observed were between both VAS scales and the appearance subscale of the GO-QOL.

[Table 9 near here]

Stage 3 Exploring discriminant validity and the relationships between subjective and objective measures of disease severity

The third and final stage was to explore the discriminant validity of the GO-QOL as well as the relationships between the subjective and objective clinical measures for TED used in this study.

Clinician ratings on BCCOM at the 48 week time point were used to split participants into two groups: condition has deteriorated (-1, N=58), or improved (+1, N=36; Table 10). Analyses were then undertaken to determine any differences in mean GO-QOL scores between these groups. It might be expected that there would be differences between the mean changes (pre/post-treatment) in GO-QOL scores depending on whether clinicians reported visual functioning/appearance having either deteriorated or improved; i.e. we would expect larger increases in scores on the GO-QOL for those

participants where clinician's reported an improvement compared to those for whom they did not.

[Table 10 near here]

Independent samples t-tests were undertaken to determine differences between the groups on both the GO-QOL and WHOQOL measures while effect sizes were calculated using Cohen's *d* (Cohen, 1988). For broad interpretation purposes threshold values for statistically significant effects for the statistic, *d* are: $0 < d < 0.1$ a trivial effect; $0.1 < d < 0.2$ a small effect; $0.2 < d < 0.5$ a moderate effect; $0.5 < d < 0.8$ a medium size effect; $0.8 < d < 1.3$ a large effect; $1.3 < d < 2.0$ a very large effect; while $d > 2.0$ is a huge effect. However, these are, at best, guidelines and the value of *d* is very much context dependent.

The independent samples t-tests showed a significant difference for the Visual Function subscale at 48-weeks (p=0.006) with a medium effect size (d=0.6; Table 11). Although no similar statistically significant difference was seen for the Appearance subscale at the same time point a moderate effect size was recorded (d=0.3). Similarly, no statistically significant differences between WHOQOL domain scores were observed for the BCCOM groups at the same time point, and effect sizes ranged from small (d=0.2 for the physical and environment domains) to moderate and medium (d=0.4 and 0.5 for the psychological and social domains, respectively).

[Table 11 near here]

The cohort was also split in relation to the scores achieved on the two Visual Analogue Scales, where 5.1 was used as the cut off to identify those of the 100 (where we had these data at week 48) who rated themselves as still having TED that was either noticeable or severe. When analysed according to these groupings, the subscale domains of the WHOQOL still showed no significant differences, but in relation to perceived noticeability of TED both subscales of the GO-QOL showed statistical significantly differences ($p<.001$) with a medium effect size for the Visual Function subscale (d=0.8) and a very large effect size for the Appearance subscale (d=1.6; Table 12). In relation to groupings based on perceived severity of TED both subscales of the

GO-QOL showed statistical significantly differences (p<.001), and again a medium effect size for the Visual Function subscale (d=0.8) and a very large effect size for the Appearance subscale (d=1.2; Table 12).

[Table 12 near here]

So it would seem that the GO-QOL does indeed demonstrate discriminant validity, although there is some suggestion in these data, that it may be dependent on the grouping variable utilised. The clinician ratings (BCCOM) and participant subjective ratings (VAS scales) were broadly similar in their results for the Visual Function component of the GO-QOL, but to distinguish in relation to appearance issues, it would seem that using participant ratings results in stronger, statistically significant differences.

Discussion

The results of this project support both the internal validity and reliability of the English translation of the GO-QOL as an outcome measure for patients with moderately active Graves' orbitopathy. Longitudinal validity has also been confirmed, with the GO-QOL being more sensitive to changes in TED over time than the more general WHOQOL-Bref., whilst also being associated with generally larger effect sizes, with both measures show an improving trend for the study participants over time. This longitudinal pattern of change for the subscales of the GO-QOL has also been shown over a 24 week period in a recent study testing Teprotumumab for thyroid-associated ophthalmology (Smith et al, 2017). While both subscales of the GO-QOL showed significant change over time in Smith et al's study, the Visual Function subscale was the one with the greatest change over time, as also suggested in our data.

The analysis with the transitional variables is more equivocal. The hypothesized changes in clinical characteristics were generally associated with significant correlations for both quality of life measures with the transitional variables at the 12 week time point. By the 48 week time point only the GO-QOL appearance subscale recorded significant correlations with the T1 (my eye condition causes me pain and discomfort) and T2 (my eye condition limits my ability to do the things I want to do).

Given the focus of the T1 and T2 variables on what might be considered to be issues more associated with visual function, it is curious that the observed correlations are with the appearance subscale. The Visual Analogue Scales in the study are more focused on what might be considered appearance issues – the perceived noticeability and severity of the TED. It is probably no surprise that these were significantly correlated with the appearance subscale of the GO-QOL at both the 12 week and 48 week study time points.

The GO-QOL has also demonstrated discriminant validity, with the visual function subscale being able to detect changes in groups identified by clinicians (using BCCOM ratings of improvement or deterioration), while both subscales could detect group differences when based on participants' ratings of TED noticeability and severity. It is worth noting that BCCOM is constructed to include factors such as diplopia which would correlate much better with function than with appearance (Appendix 1). It has been suggested that patients with TED overrated the extent to which their appearance was affected, while endocrinologists underrated it (Terwee et al, 2003). Of course, the measures employed do not necessarily take into account the reference point, i.e. patients are possibly comparing themselves prior to TED, or with other people who do not have the disease at all, whereas clinicians could be comparing across individuals who have TED. Without interviewing the participants and clinicians concerned it is impossible to know.

Historically health care professionals have found it difficult to engage with patients' concerns regarding their looks and body image (Bessell et al, 2012), quantifying these concerns with a measure like the GO-QOL may help doctors' engagement with their patients concerns and target therapy accordingly.

Competing interest statement: The authors of this study have no competing interests to declare.

Acknowledgements:
The authors would like to thank the staff at all the CIRTED trial sites for their assistance in this project.

The authors would like to thank the staff at the University of the West of England who have contributed to the study design, data collection and data management; in particular, Jane Murray, Nicky Rumsey, Emma Williams & Laura Kingston.

The material in this publication is the result of use of the WHOQOL-UK and the assistance of the University of Bath and the World Health Organisation is acknowledged.

RL received support from the National Institute for Health Research (NIHR) Biomedical Research Centre based at Moorfields Eye Hospital NHS Foundation Trust and University College London Institute of Ophthalmology. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR or the Department of Health

List of CIRTED Trial Investigators:

Moorfields Eye Hospital NHS Foundation Trust, London, UK: Rathie Rajendram, Nicola Harris, Olivia C Morris, Catey Bunce, Daniel Ezra, Geoff Rose
Bristol Eye Hospital, University Hospitals Bristol NHS Foundation Trust: Marjorie Tomlinson, Sue Yarrow, Helen Garrott, Helen Herbert, Andrew Dick, Mike Potts
Manchester Royal Eye Hospital, Central Manchester NHS Foundation Trust: Anne Cook
Wade Centre for Radiotherapy Research, The Christie NHS Foundation Trust: Rao Gattamaneni
Western Eye Hospital, Imperial College NHS Healthcare Trust: Rajni Jain, Jane Olver
University College London Hospitals NHS Foundation Trust: Steven Hurel, Fion Bremner
Tennent Institute of Ophthalmology, Gartnavel General Hospital, NHS Greater Glasgow and Clyde: Suzannah R Drummond, Ewan Kemp
Beatson West of Scotland Cancer Centre: Diana Ritchie
University Hospital of Wales: Daniel Morris, Carol Lane
Thyroid Research Group, Cardiff University School of Medicine, Chunhei Li, Julie Pell, Robert Hills
Velindre NHS Trust, Velindre Cancer Centre: Nachi Palaniappan

Funders:

Above and Beyond Charities, National Eye Research Centre, Moorfields Eye Hospital Special Trustees.

Sponsor:

University of Bristol

References

Bahn, R. S., & Heufelder, M. D. (1993). Pathogenesis of Graves' Ophthalmopathy. *N Engl J Med*, 329(20), 1448-75.

Bahn, R. S., & Gorman, C. A. (1987). Choice of therapy and criteria for assessing treatment outcome in thyroid-associated ophthalmopathy· *Endocrinology and metabolism clinics of North America,* 16(2), 391-407.

Bartalena, L., Baldeshi, L., Boboridis, K., Eckstein, A., Kahaly, G. J., Marcocci, C., … Wiersinga, W. M. (2016). The 2016 European Thyroid Association / European Group on Graves ' Orbitopathy Guidelines for the Management of Graves ' Orbitopathy. *European Thyroid Journal*, 5, 9–26. doi: https://doi.org/10.1159/000443828

Bessell, A., Dures, E., Semple, C., & Jackson, S. (2012). Addressing appearance related distress across clinical conditions. *British Journal of Nursing*, 21(19), 1138–1144.

British Psychological Society Steering Committee on Test Standards. (1992). *Psychological testing: a guide.* Leicester: British Psychological Society.

Carr, T. (1997). Assessment and measurement in clinical practice. R. Lansdown, N. Rumsey, E. Bradbury, T. Carr, & J. Partridge (Eds.), *Visibly Different.* Oxford: Butterworth-Heinemann.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences.* USA: Lawrence Erlbaum Associates.

Coulter, I., Frewin, S., Krassas, G. E., & Perros, P. (2007). Psychological implications of Graves' orbitopathy. *European Journal of Endocrinology*, 157(2), 127–131. doi: 10.1530/EJE-07-0205

Dickinson, A. J., & Perros, P. (2001). Controversies in the clinical evaluation of active thyroid-associated orbitopathy: use of a detailed protocol with comparative photographs for objective assessment· *Clin Endocrinol,* 55(3), 283-303.

Haggerty, H., Richardson, S., Mitchell, K. W. Dickinson, A. J. (2005). A modified method for measuring uniocular fields of fixation: reliability in healthy subjects and in patients with Graves orbitopathy· *Archives of ophthalmology,* 123(3), 356-62.

Khoo, T. K., & Bahn, R. S. (2007). Pathogenesis of Graves' ophthalmopathy: the role of autoantibodies. *Thyroid,* 17(10), 1013-8.

Marcocci, C., Bartalena, L., Tanda, M. L., Manetti, L., Dell-Unto, E., Rocchi, R., … Pinchera, A. (2001). Comparison of the effectiveness and tolerability of intravenous or oral glucocorticoids associated with orbital radiotherapy in the management of severe Graves' ophthalmopathy: results of a prospective, single-blind, randomized study. *Journal of Clininical Endocrinology & Metabolism,* 86(8), 3562-7. doi: http://dx.doi.org/10.1210/jcem.86.8.7737

Marcocci, C., Kahaly, G. J., Krassas, G. E., Bartalena, L., Prummel, M., Stahl, M., … Wiersinga, W. M. (2011). Selenium and the course of mild Graves' orbitopathy. *New England Journal of Medicine*, 364, 20, 1920–1931. doi: 10.1056/NEJMoa1012985

Mourits, M. P., Prummel, M. F., Wiersinga, W. M., & Koornneef, L. (1997). Clinical activity score as a guide in the management of patients with Graves' ophthalmopathy. *Clin Endocrinol (Oxf)* 1997; 47(1): 9-14.

Mourits, M. P., van Kempen-Harteveld, M. L., Garcia, M. B., Koppeschaar, H. P., Tick, L., & Terwee, C. B. (2000). Radiotherapy for Graves' orbitopathy: randomised placebo-controlled study. *Lancet* 2000; 355(9214): 1505-9.

Perros, P., Crombie, A. L., & Kendall-Taylor, P. (1995). Natural history of thyroid associated ophthalmopathy. *Clinical Endocrinology*, 42(1), 45–50.

Perros, P., Crombie, A. L., Matthews, J. N., & Kendall-Taylor, P. (1993). Age and gender influence the severity of thyroid-associated ophthalmopathy: a study of 101 patients attending a combined thyroid-eye clinic. *Clin Endocrinol (Oxf),* **38**(4): 367-72.

Prummel, M. F., Terwee, C. B., Gerding, M.N., Baldeschi, L., Mourits, M. P., Blank, L., … Wiersinga, W. M. (2004). A randomized controlled trial of orbital radiotherapy versus sham irradiation in patients with mild Graves' ophthalmopathy. *Journal of Clininical Endocrinology & Metabolism,* 89(1), 15-20. doi: https://doi.org/10.1210/jc.2003-030809

Rajendram, R., Lee, R. W. J., Potts, M. J., Rose, G. E., Jain, R., Olver, J. M., … Uddin, J. (2008). Protocol for the combined immunosuppression and radiotherapy in thyroid eye disease (CIRTED) trial: A multi-centre, double-masked, factorial randomised controlled trial. *Trials*, 9(6), 1–17.

Skevington, S. M., Lotfy, M., & O'Connell, K. A. (2004). The World Health Orgaization's WHOQOL-Bref quality of life assessment: Psychometric properties and results of the international field trial. A report from the WHOQOL Group. *Quality of Life Research*, 13, 299-310.

Smith, T. J., Kahaly, G. J., Ezra, D. G., Fleming, J. C., Dailey, R. A., Tang, R. A., … Douglas, R. S. D. (2017). Teprotumumab for Thyroid-Associated Ophthalmopathy. New England Journal of Medicine, 376:1748-61. doi: 10.1056/NEJMoa1614949

Terwee, C. B., Dekker, F. W., Bonsel, G. J., Heisterkamp, S. H., Prummel, M. F., Baldeschi, L., & Wiersinga, W.M. (2003). Facial disfigurement: is it in the eye of the beholder? A study in patients with Graves' ophthalmopathy. *Clinical Endocrinology,* 58, 192–198.

Terwee, C. B., Gerding, M. N., Dekker, F. W., Prummel, M. F., van der Pol, J. P., & Wiersinga, W. M. (1999). Test-retest reliability of the GO-QOL : A disease-specific quality of life questionnaire for patients with Graves' Ophthalmopathy. *Journal of Clinical Epidemiology*, 52(9), 875–884. doi: https://doi.org/10.1016/S0895-4356(99)00069-4

Terwee, C. B., Gerding, M., Dekker, F., Prummel, M., & Wiersinga, W. M. (1998). Development of a disease specific quality of life questionnaire for patients with Graves' ophthalmopathy: the GO-QOL. *British Journal of Ophthalmology*, 82(7), 773–779.

Weetman, A. P. (1991). Thyroid-associated eye disease: pathophysiology.

Lancet, 338(8758), 25–8.

WHOQOL Group, The. (1998). Development of the World Health Organisation WHOQOL-BREF Quality of Life Assessment. *Psychological Medicine,* 28 (3), 551-558.

APPENDICES

Appendix 1

Binary composite clinical outcome measure (BCCOM)

The BCCOM is a binary outcome with a positive result with no deterioration vs no change or deterioration in any component of the score.

Major Criteria

- An improvement of ≥ 1 grade in diplopia score*
- An improvement of >8 degrees of eye movement in any direction#
- A reduction of ≥ 2 mm in proptosis

Minor Criteria

- A reduction of ≥ 2 mm in lid aperture
- An improvement of ≥ 1 grade in soft tissue involvement†
- An improvement in best-corrected visual acuity of ≥ 1 line on the Snellen chart·
- Subjective improvement^

Response to Treatment

Very good: ≥ 2 major criteria
Good: 1 major criterion
Fair: ≥ 2 minor criteria
No Change: 1 minor criterion
Worse: Deterioration in at least 1 major or 2 minor criteria

* Bahn, R. S., & Gorman, C. A. (1987). Choice of therapy and criteria for assessing treatment outcome in thyroid-associated ophthalmopathy· *Endocrinology and metabolism clinics of North America,* 16(2), 391-407.

Haggerty, H., Richardson, S., Mitchell, K. W. Dickinson, A. J. (2005). A modified method for measuring uniocular fields of fixation: reliability in healthy subjects and in patients with Graves orbitopathy· *Archives of ophthalmology,* 123(3), 356-62.

† Dickinson, A. J., & Perros, P. (2001). Controversies in the clinical evaluation of active thyroid-associated orbitopathy: use of a detailed protocol with comparative photographs for objective assessment· *Clin Endocrinol,* 55(3), 283-303.

^ Marcocci, C., Bartalena, L., Tanda, M. L., Manetti, L., Dell-Unto, E., Rocchi, R., … Pinchera, A. (2001). Comparison of the effectiveness and tolerability of intravenous or oral glucocorticoids associated with orbital radiotherapy in the management of severe Graves' ophthalmopathy: results of a prospective, single-blind, randomized study. *Journal of Clininical Endocrinology & Metabolism,* 86(8), 3562-7. doi: http://dx.doi.org/10.1210/jcem.86.8.7737

Table 1. Overview of the standardised questionnaires.

Measure	**Description**
Standardised questionnaires	
GO-QOL	- TED specific quality of life measure, validated in the Netherlands - 2 subscales: 'visual function' & 'appearance' comprising 8 questions each - each item is scored as follows: 1= not impaired; 2=a little impaired; 3=severely impaired - raw scores are transformed to give a total out of 100 for each subscale - higher scores indicate greater quality of life
WHOQOL-Bref	- general quality of life measure, validated for use in the UK - 4 subscales: psychological, physical, social and environmental; 28 questions in total - scored on a 5-point Likert scale - raw scores are transformed to give a total out of 100 for each subscale - higher scores indicate greater satisfaction with life
Transitional variables	
T1	- is a single item: "My eye condition causes me physical pain/discomfort" - scored as follows: 1= never/almost never; 2=sometimes; 3=often; 4=almost always
T2	- is a single item: "My eye condition limits my physical ability to do the things I want to do" - each item is scored as follows: 1= never/almost never; 2=sometimes; 3=often; 4=almost always
Visual Analogue Scales	

Noticeability	- is a single item: "How noticeable do you feel your thyroid eye disease is to other people?" - scored on a 10 cm line with the following anchors: Not at all noticeable |.....| Very noticeable - scored 0-10 - higher scores indicate greater distress
Severity	is a single item: "How severe do you feel your thyroid eye disease is?" - scored on a 10 cm line with the following anchors: Not very severe |.....| Extremely severe - scored 0-10 - higher scores indicate greater distress
Clinical rating	
Binary Composite Clinical Outcome Measure (BCCOM)	- clinician-rating - binary composite outcome score with a positive result (improvement with no concomitant deterioration) versus no change or any deterioration (see Appendix 1) - deteriorated (-1), improved (1)

Table 2. Data collection time points and their relation to participant trial appointments, plus study measures used at each time point, and number of potential participants attending each appointment (N)

		Trial Appointment	Study measures	N
Study time points	**1**	-2 weeks Enrolment	GO-QOL, WHOQOL, VAS, demographic data	167
	2	0 weeks Randomization	GO-QOL	133
	3	12 weeks Short term	GO-QOL, WHOQOL, transition variables, VAS	108
	4	48 weeks Long term	GO-QOL, WHOQOL, transition variables, VAS, BCCOM	102

Table 3. Participant characteristics for Stage 1 GO-QOL Validation: internal validity and test re-test

	Study time point	
Variable	**Enrolment (-2 weeks)**	**Randomization (week 0)**
Sample size	142	126
Age	48.02±11.44	48.27±11.40
Sex (male/female)	41/101	34/92
Ethnicity		
Caucasian/Black	99/16	87/14
Asian/Oriental	11/4	10/4
Other (or not stated)	12	11

Table 4. Item response table for GO-QOL at time point 1 (-2 weeks, enrolment).

	N	Min score	Max score	Mean score	SD
Visual Function subscale					
VF1 Cycling	87	1	3	2.43	0.80
VF2 Driving	103	1	3	2.19	0.81
VF3 Walking indoors	139	1	3	2.71	0.54
VF4 Walking outdoors	141	1	3	2.49	0.64
VF5 Reading	140	1	3	2.01	0.70
VF6 Watching TV	141	1	3	2.12	0.68
VF7 Hobbies	122	1	3	2.29	0.77
VF8 Interference with daily life	141	1	3	2.10	0.77
Appearance subscale					
App9 Change in appearance	142	1	3	1.42	0.54
App10 Feeling watched	142	1	3	2.14	0.77
App11 Unpleasant reactions	140	1	3	2.49	0.68
App12 Impact on self-confidence	142	1	3	1.73	0.72
App13 Feeling of social isolation	142	1	3	2.50	0.69
App14 Influence on friendships	142	1	3	2.46	0.73
App15 Less often in photos	141	1	3	1.90	0.85
App16 Camouflaging appearance	142	1	3	2.04	0.82

Key: N= number of responses; SD = standard deviation

Table 5. Item response table for GO-QOL at time point 2 (week 0, randomisation)

	N	Min Score	Max Score	Mean score	SD
Visual Function subscale					
VF1 Cycling	75	1	3	2.59	0.68
VF2 Driving	93	1	3	2.38	0.75
VF3 Walking indoors	125	1	3	2.72	0.47
VF4 Walking outdoors	126	1	3	2.61	0.61
VF5 Reading	126	1	3	2.23	0.69
VF6 Watching TV	126	1	3	2.36	0.66
VF7 Hobbies	110	1	3	2.47	0.69
VF8 Interference with daily life	126	1	3	2.28	0.73
Appearance subscale					
App9 Change in appearance	126	1	2	1.50	0.50
App10 Feeling watched	126	1	3	2.17	0.76
App11 Unpleasant reactions	125	1	3	2.58	0.60
App12 Impact on self-confidence	126	1	3	1.83	0.75
App13 Feeling of social isolation	126	1	3	2.59	0.65
App14 Influence on friendships	126	1	3	2.45	0.77
App15 Less often in photos	126	1	3	2.01	0.83
App16 Camouflaging appearance	126	1	3	2.05	0.81

Key: N= number of responses; SD = standard deviation

Table 6. CIRTED trial allocation groups.

Group	Allocation	N=126 randomised
1	Radiotherapy and Azathioprine	31 (24.6%)
2	SHAM Radiotherapy and Azathioprine	31 (24.6%)
3	Radiotherapy and PLACEBO	32 (25.3%)
4	SHAM Radiotherapy and PLACEBO	32 (25.3%)

Table 7. Participant characteristics for Stage 2: Longitudinal validation of GO-QOL

Variable	**Study time point**	
	Short term (12 weeks)	**Long term (48 weeks)**
N	108	101
Age	49.48±10.82	49.88±10.46
Sex (male/female)	29/79	27/73
Ethnicity		
Caucasian/Black	79/13	72/10
Asian/Oriental	8/4	9/4
Other (or not stated)	4	5

Table 8: Mean scores ± standard deviations for G0-QOL and WHOQOL-Bref across study time points, results for RM ANOVA with effect sizes

Study measure	*Enrolment (-2 weeks)*	*Short term (12 weeks)*	*Long term (48 weeks)*	*RM ANOVA F & p value*	*Effect size* η^2
GO-QOL subscales:					
Visual Function	65.94 ±28.49	72.49 ±26.64	76.11 ±24.73	11.034 p<0.001	0.114 Medium
Appearance	54.1 ±26.54	58.81 ±27.25	60.84 ±28.75	13.061 p=0.002	0.069 Small
WHOQOL-Bref subscales:					
Physical	58.69 ±22.64	60.93 ±21.01	63.46 ±20.67	3.18 p=0.044	0.037 Small
Psychological	52.66 ±21.35	52.14 ±23.04	57.79 ±20.20	6.53 p=0.002	0.064 Small
Environment	66.60 ±20.03	67.81 ±18.92	69.15 ±17.13	1.97 p=0.143	0.043 Small
Social Relationships	66.08 ±20.81	60.83 ±21.53	65.54 ±21.77	2.99 p=0.054	0.041 Small

Table 9. Transition scores and visual analogue scales correlated with GO-QOL and WHOQOL scores.

	T1	T2	VAS Noticeability	VAS Severity
Week 12 GO-QOL subscales:				
Visual Function	-0.413**	0.015	-0.108	-0.117
Appearance	-0.409**	-0.580**	-0.587**	-0.552**
Week 12 WHOQOL-Bref subscales:				
Psychological	-0.284**	-0.475**	-0.399**	-0.127
Physical	-0.451**	-0.612**	-0.383**	-0.466**
Social	-0.193	-0.307**	-0.167	-0.262*
Environment	-0.295**	-0.508**	-0.346**	-0.253**
Week 48 GO-QOL subscales:				
Visual Function	-0.017	-0.162	0.014	-0.046
Appearance	-0.389**	-0.574**	-0.723**	-0.678**
Week 48 WHOQOL-Bref subscales:				
Psychological	0.19	0.017	-0.037	0.148
Physical	0.149	-0.005	-0.106	-0.024
Social	0.158	0.083	-0.151	0.089
Environment	0.199	0.025	-0.07	0.006

Key: ** = significant at p<.001, *= significant at p<.05

Table 10. Participant characteristics for Stage 3: Exploring discriminant validity based on BCCOM

Variable	**Study time point 48 weeks**	
	-1 deteriorated	**+1 improved**
N	58	36
Age	50.31±10.72	48.72±10.53
Sex (male/female)	12/46	12/24
Ethnicity		
Caucasian/Black	46/4	22/6
Asian/Oriental	5/2	4/1
Other/missing	1/0	3/0

Table 11. Comparing means of GO-QOL & WHOQOL subscales according to BCCOM group (improved n=36, deteriorated n=58) at 48 week time point

GO-QOL	**BCCOM group**	**Mean±sd**	**p value**	**Cohen's d (CI)**
Visual Function	+1 (improved)	83.06±21.06	0.006	0.6 medium
	-1(deteriorated)	67.86±26.46		(CI 0.19 – 1.04)
Appearance	+1 (improved)	64.89±28.33	0.152	0.3 moderate
	-1 (deteriorated)	55.89±28.26		(CI -0.13 – 0.73)
WHOQOL				
Psychological	+1 (improved)	51.73±22.32	0.138	0.4 moderate
	-1(deteriorated)	59.61±19.29		(CI -0.03 – 0.83)
Physical	+1 (improved)	58.08±22.27	0.537	0.2 small
	-1(deteriorated)	61.63±22.21		(CI -0.23 – 0.63)
Social	+1 (improved)	56.30±23.75	0.083	0.5 medium
	-1(deteriorated)	67.34±22.96		(CI -0.89 – -0.05)
Environment	+1 (improved)	65.08±19.66	0.356	0.2 small
	-1(deteriorated)	69.74±19.68		(CI -0.23 – 0.63)

Table 12. Comparing means of GO-QOL subscales according to VAS group at 48 week time point

GO-QOL	VAS noticeability	Mean±sd	p value	Cohen's d (CI)
Visual Function	<5.0 (improved, n=45)	84.63±20.72	0.001	0.8 medium
	>5.1(noticeable, n=54)	66.37±26.36		(CI 0.41 – 1.19)
Appearance	<5.0 (improved, n=45)	79.86±18.17	0.001	1.6 very large
	>5.1(noticeable, n=55)	45.23±25.11		(CI 1.17 – 2.03)
	VAS severity			
Visual Function	<5.0 (improved, n=53)	84.00±20.46	0.001	0.8 medium
	>5.1(severe, n=45)	64.54±26.71		(CI 0.41 – 1.19)
Appearance	<5.0 (improved, n=53)	74.29±23.14	0.001	1.2 large
	>5.1(severe, n=46)	46.06±25.60		(CI 0.77 – 1.63)