BMJ Open

Psychometric evaluation of the NTDT-PRO questionnaire for assessing symptoms in patients with non-transfusion-dependent beta-thalassaemia

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-066683
Article Type:	Original research
Date Submitted by the Author:	15-Jul-2022
Complete List of Authors:	Taher, Ali T.; American University of Beirut Medical Center, Department of Internal Medicine Musallam, Khaled M.; Thalassemia Center; International Network of Hematology Viprakasit, Vip; Mahidol University, Division of Hematology & Oncology, Department of Pediatrics & Siriraj Thalassemia Center, Siriraj Research Hospital Kattamis, Antonis; National and Kapodistrian University of Athens, First Department of Pediatrics Lord-Bessen, Jennifer; Bristol Myers Squibb Co Yucel, Aylin; Bristol Myers Squibb Co Guo, Shien; Evidera Waltham Pelligra, Christopher; Evidera Shields, Alan L.; Adelphi Values Boston Shetty, Jeevan K.; Celgene International Sàrl Miteva, Dimana; Celgene International Sàrl Bueno, Luciana Moro; Celgene International Sàrl Cappellini , MD; University of Milan, Department of Internal Medicine, Fondazione IRCCS Ca' Granda Policlinico Hospital
Keywords:	Anaemia < HAEMATOLOGY, Blood bank & transfusion medicine < HAEMATOLOGY, Clinical trials < THERAPEUTICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Psychometric evaluation of the NTDT-PRO questionnaire for assessing symptoms in patients with non-transfusion-dependent beta-thalassaemia

Ali T. Taher,[1] Khaled M. Musallam,[2,3] Vip Viprakasit,[4] Antonis Kattamis,[5] Jennifer Lord-Bessen,[6] Aylin Yucel,[6] Shien Guo,[7] Christopher Pelligra,[8] Alan L. Shields,[9] Jeevan K. Shetty,[10] Dimana Miteva,[10] Luciana Moro Bueno,[10] Maria Domenica Cappellini[11]

[1]Department of Internal Medicine, American University of Beirut Medical Center, Beirut, Lebanon

[2]Thalassemia Center, Burjeel Medical City, Abu Dhabi, UAE

[3]International Network of Hematology, London, UK

[4]Division of Hematology & Oncology, Department of Pediatrics & Siriraj Thalassemia Center, Siriraj Research Hospital, Mahidol University, Bangkok, Thailand

[5]First Department of Pediatrics, National and Kapodistrian University of Athens, Athens, Greece

[6]Bristol Myers Squibb, Princeton, NJ, USA

[7]Evidera, Waltham, MA, USA

[8]Evidera, Bogotá, Colombia

[9]Adelphi Values, Boston, MA, USA

[10]Celgene International Sàrl, a Bristol-Myers Squibb Company, Boudry, Switzerland

[11]Department of Internal Medicine, Fondazione IRCCS Ca' Granda Policlinico Hospital, University of Milan, Milan, Italy

ORCID

AT Taher: 0000-0001-8515-2238

KM Musallam: 0000-0003-3935-903X

V Viprakasit: 0000-0003-3162-1849

A Kattamis: 0000-0002-5178-0655

C Pelligra: 0000-0002-5255-2777

MD Cappellini: 0000-0001-8676-6864

Correspondence

Ali T. Taher, M.D., Ph.D., F.R.C.P.

American University of Beirut Medical Center, Halim and Aida Daniel Academic and Clinical Center, 4th floor, Hamra, Beirut, Lebanon

Telephone: +9611350000 Extension 5392

Email: ataher@aub.edu.lb

Keywords (3-6 words/phrases): psychometrics; non-transfusion-dependent beta-thalassaemia; patient-reported outcomes; symptom; anaemia

Running title: NTDT-PRO psychometric evaluation

Abstract

Objectives The NTDT-PRO questionnaire was developed for assessing anaemia-related Tiredness/Weakness (T/W) and Shortness of Breath (SoB) among patients with non-transfusion-dependent β-thalassaemia (NTDT). Its psychometric properties were evaluated in this study using data from the BEYOND trial (NCT03342404).

Design A retrospective study.

Methods Participants (N=145) completed the NTDT-PRO daily from baseline until week 24, and the 36-Item Short Form Health Survey version 2 (SF-36v2®), Functional Assessment of Chronic Illness Therapy – Fatigue (FACIT-F), and Patient Global Impression of Severity (PGI-S) at select time points.

Results Cronbach's alpha at weeks 13–24 was 0.95 and 0.84 for the T/W and SoB domains, respectively, indicating acceptable internal consistency reliability. Among participants self-reporting no change in thalassaemia symptoms via the PGI-S between baseline and week 1, intraclass correlation coefficients were 0.94 and 0.92 for the T/W and SoB domains, respectively, indicating excellent test–retest reliability. In a known-groups validity analysis, least-squares mean T/W and SoB scores at weeks 13–24 were worse in participants with worse scores for the FACIT-F Fatigue Subscale (FS), SF-36v2® vitality, or PGI-S. Indicating responsiveness, changes in T/W and SoB domain scores were moderately correlated with changes in haemoglobin levels, and strongly correlated with changes in SF-36v2® vitality, FACIT-F FS, select FACIT-F items, and the PGI-S. Improvements in least-squares mean T/W and SoB scores were higher in participants with greater improvements in scores on other patient-reported outcomes measuring similar constructs.

Conclusion The NTDT-PRO demonstrated adequate psychometric properties to assess anaemia-related symptoms in adults with NTDT and can be used to evaluate treatment efficacy in clinical trials.

Strengths and limitations of this study

- Strengths of this study include use of well-validated PRO instruments such as PGI-S, PGI-C, SF-36v2®, and FACIT-F.
- The data used in this analysis were from a phase 2 interventional study with participants from multiple geographic regions and spanning a range of NTDT symptom severities.
- The use of data from an interventional study allowed for changes in symptom severity to be observed, validating NTDT-PRO's sensitivity to identify longitudinal changes in symptoms.
- Given that NTDT is a rare disease, limitations of the present study include the reduced sample size for typical psychometric evaluations.
- Cut-off values used to define different levels of improvement in the responsiveness analysis are not well established and were based on certain assumptions.

INTRODUCTION

β-thalassaemias are a group of genetic blood disorders characterised by defective synthesis of the β-globin chains of haemoglobin and ineffective erythropoiesis. Phenotypes are highly variable: while some patients are borderline asymptomatic, others experience significant symptoms associated with severe chronic anaemia.[1]

From a clinical perspective, patients are often categorised as having transfusion-dependent β-thalassaemia (TDT) or non-transfusion-dependent β-thalassaemia (NTDT). While patients with TDT require lifelong blood transfusions, those with NTDT only require transfusions in certain circumstances, such as during infections, pregnancy, and surgery.[2,3] Due to anaemia or primary iron overload, which accumulate as patients get older, NTDT can result in various comorbidities (e.g., hepatic disease, endocrinopathy, thromboembolic events, pulmonary hypertension, leg ulcers, and extramedullary haematopoietic [EMH] masses), which not only have a negative impact patients' daily activities and quality of life (QoL), but also reduces survival.[4-6]

Patient-reported outcomes (PRO) questionnaires are used to assess how patients feel and function as well as their overall QoL. Reflecting the patient experience in these ways is important when evaluating treatments in clinical trials, and particularly in instances when patients experience symptoms from lifelong diseases.

Patient-centred research in NTDT is limited by a lack of rigorously developed PRO instruments for assessing symptoms important to patients in the target patient population. For example, health-related QoL (HRQoL) in patients with β-thalassaemias has typically been evaluated by generic questionnaires such as the Short Form Health Survey version 2 (SF-36v2®) and the World Health Organization 100-item Quality of Life Survey (WHOQOL-100),[7,8] which may fail to capture the unique experiences of patients with β-thalassaemia. Two β-thalassaemia-specific PRO instruments for assessing HRQoL are now available: the Specific Thalassaemia Quality of Life Instrument (STQOLI) and the Transfusion-dependent Quality of Life (TranQoL) questionnaire.[9,10] However, both tools were developed for patients with TDT and include questions on the impact of transfusions, which are often not relevant for patients with NTDT. Moreover, they focus more on general functioning and

QoL and do not specifically capture anaemia-related symptoms of β-thalassaemias, which can be more prominent in NTDT than in TDT because of the lack of transfusions.[11,12] In addition, neither instrument has been evaluated in patients with NTDT.

The NTDT-PRO was created to fill the gap in available, indication-specific PRO questionnaires defensible for use among patients with NTDT. Developed in the context of evaluating the treatment benefit of luspatercept (an approved treatment for anaemia in adults with TDT) among patients with NTDT, the NTDT-PRO is a 6-item questionnaire intended to measure the most relevant and important anaemia-related symptoms of NTDT.[13] In accordance with US Food and Drug Administration (FDA) guidance on the development of PRO tools,[14] evidence supporting the content validity of the NTDT-PRO was obtained from qualitative work, including concept elicitation and cognitive interviews with patients with NTDT,[13] and a preliminary psychometric evaluation using data from a 24-week observational study showed promising reliability and validity results.[15] However, the ability of the NTDT-PRO to capture longitudinal changes in symptoms could not be properly assessed due to the non-interventional study design. In the present study, a detailed evaluation of the reliability and validity of the NTDT-PRO was conducted, including its ability to reflect changes in symptom severity over time, using data from the BEYOND trial [16])

METHODS

Study design

The analysis was based on data generated from BEYOND, a phase 2, double-blind, randomised, placebo-controlled trial of luspatercept in adults with NTDT (NCT03342404), conducted in the USA, Greece, Italy, Lebanon, Thailand, and the UK [16]) Briefly, the trial included double-blind and open-label treatment phases and long-term follow-up. For double-blind treatment, participants were randomly assigned 2:1 to luspatercept or placebo. Luspatercept was administered as a subcutaneous injection every 3 weeks for 48 weeks. The assessment period for the primary and key secondary efficacy endpoints was weeks 13–24. The starting dose of luspatercept was 1 mg/kg and the maximum dose was 1.25 mg/kg or 120 mg. The trial was unblinded 48 weeks after the last participant had received their first dose of study drug. All participants were eligible to receive open-label

luspatercept for up to 15 months, and could then continue to receive luspatercept during the post-treatment follow-up period.

BEYOND received institutional review board/ethics committee approval and was conducted in accordance with International Council for Harmonisation Good Clinical Practice and the Declaration of Helsinki.

Participants

Participants were adults (≥18 years of age) with β-thalassaemia or haemoglobin E/β-thalassaemia. They were non-transfusion-dependent, as defined by receipt of 0 to 5 units of red blood cells during the 24 weeks before randomisation, and had not received a red blood cell transfusion in the 8 weeks prior to randomisation. To be eligible for enrolment, they were additionally required to have a mean baseline haemoglobin level (based on at least 2 measurements taken ≥1 week apart) of ≤10.0 g/dL and an Eastern Cooperative Oncology Group (ECOG) performance status of 0 or 1. Patients with haemoglobin S/β-thalassaemia or α-thalassaemia alone were excluded, as were patients who had previously been exposed to luspatercept or sotatercept. All participants provided written informed consent.

Patient and public involvement

None.

PRO assessments

The NTDT-PRO and Patient Global Impression of Severity (PGI-S) were administered daily from the 7 days prior to randomisation until week 24, then daily for 7 days before dosing of every other dose of study drug. The Patient Global Impression of Change (PGI-C), SF-36v2®, and Functional Assessment of Chronic Illness Therapy – Fatigue (FACIT-F) were administered at screening and on the day of dosing for every other dose of study drug, starting from the first dose. The SF-36v2®, FACIT-F, and PGI-C assessments were mapped to a nominal week using a mapping algorithm (see online supplementary table S1).

NTDT-PRO

NTDT-PRO assesses the severity of symptoms associated with NTDT in the 24 hours prior to administration. The 6 items assess tiredness (lack of energy, 2 items), weakness (lack of strength, 2 items), and shortness of breath (2 items) when doing and when not doing physical activity. Each item uses an 11-point numeric rating scale (NRS) ranging from 0 (no symptom) to 10 (extreme symptom). Responses to the NTDT-PRO can be used to derive Tiredness/Weakness (T/W) and Shortness of Breath (SoB) domain scores. In the BEYOND trial, the NTDT-PRO was completed in the evening as a part of an electronic diary that also included the PGI-S. NTDT-PRO T/W and SoB scores were included as secondary endpoints in the trial [16]).

Weekly item and domain scores were calculated from baseline (week 0) to week 24. For a given week, the weekly score for each item was calculated as the average of the daily scores for that item if scores were available for at least 4 days (i.e., at least 50% of the week); otherwise, the score was set to "missing." Weekly T/W and SoB domain scores (range: 0 [no symptoms] to 10 [extreme symptoms]) were calculated as the average of non-missing weekly item scores for the tiredness and weakness items (T/W domain) or shortness of breath items (SoB domain). Weekly domain scores were only calculated if weekly scores were non-missing for at least 2 of the 4 tiredness/weakness items (including ≥1 tiredness item and ≥1 weakness item) or at least 1 of the 2 shortness of breath items; otherwise, they were set to "missing." Average T/W and SoB scores over weeks 13–24 were calculated using data for all non-missing weeks during that time interval. If all weekly scores over weeks 13–24 were missing, the average score over weeks 13–24 was set to "missing".

PGI-S

PGI-S is a single-item questionnaire that assesses a patient's perception of their overall thalassaemia symptom severity in the previous 24 hours on an 11-point NRS ranging from 0 (no symptoms) to 10 (very severe symptoms). The weekly PGI-S score for a given week was calculated as the average of the daily scores if scores were available for at least 4 days; otherwise, it was set to "missing". Average PGI-S scores over weeks 13–24 were calculated using data for all non-missing weeks.

PGI-C

PGI-C is a single-item questionnaire that assesses a patient's perception of how their symptoms have changed over time. In BEYOND, participants responded to the question "How would you rate the overall change in your thalassaemia symptoms since the start of this study?" by selecting 1 of 7 response options ranging from "A great deal better" to "A great deal worse".

SF-36v2®

SF-36v2® consists of 8 multi-item scales assessing the following aspects of health over the previous 7 days: physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional, and mental health. SF-36v2® data were scored using Health Outcomes™ Scoring Software 5 (QualityMetric, Lincoln, RI, USA).[17] For each multi-item scale, the average of all items within the scale was calculated and the raw scores were converted to a 0 to 100 scale. They were then transformed to a US norm-based T-score (mean: 50, standard deviation [SD]: 10), with a higher T-score indicating better health. Finally, the Physical Component Summary (PCS) and Mental Component Summary (MCS) were derived as weighted averages of the T-scores for the 8 multi-item scales.

FACIT-F

FACIT-F is a 40-item questionnaire assessing fatigue and its effects on functioning and daily activities. It consists of the 27-item Functional Assessment of Cancer Therapy – General (FACT-G) questionnaire and the 13-item Fatigue Subscale (FS). All items have a 7-day recall period and are rated on a 5-point scale ranging from "Not at all" to "Very much".

FACT-G comprises 4 domains: physical well-being (7 items, range: 0 to 28 points), social/family well-being (7 items, range: 0 to 28 points), emotional well-being (6 items, range: 0 to 24 points), and functional well-being (7 items, range: 0 to 28 points). Scores for each FACT-G domain and the FS (range: 0 to 52 points) were derived by summing the scores for the individual items (after reverse scoring, as applicable).[18]

Scores for 3 additional summary scales were also calculated: FACT-G total score=sum of scores for all FACT-G items (range: 0 to 108 points); FACIT-F trial outcome index (TOI)=sum of the scores for FACT-G physical well-being, FACT-G functional well-being, and the FS (range: 0 to 108 points); and FACIT-F total score=sum of scores for all FACT-G items and the FS (range: 0 to 160 points). For the FACT-G domains, the FS, and the additional summary scales, a higher score indicates less fatigue or better HRQoL.

Statistical analyses

All statistical analyses were conducted using SAS version 9.4 (SAS Institute Inc., Cary, NC, USA). Analyses were performed on blinded data collected up to week 24 during double-blind treatment (data cut-off: January 7, 2020) using the intent-to-treat (ITT) population: all randomised participants. Summary statistics were calculated for demographics, baseline clinical characteristics, and PRO scores. For NTDT-PRO scores, floor and ceiling effects were also assessed.

Quality of completion of the NTDT-PRO was evaluated by calculating the percentages of participants with missing and non-missing weekly scores from among participants who were eligible for the assessment. Item–item and item–domain correlations for the NTDT-PRO were assessed by calculating Spearman's rank correlation coefficients, which were interpreted as <0.3=weak, ≥0.3 to <0.7=moderate, ≥0.7 to <0.9=strong, and ≥0.9=very strong.[19]

Confirmation of the weekly scoring rule

To evaluate whether modifying the weekly scoring rule for the NTDT-PRO would impact the variability of weekly item scores, an analysis was conducted at baseline, weeks 1, 2, 4, 8, 12, 16, 20, and 24, including data only from those participants with no missing daily item scores within each week. For each participant, a weekly score for each item was generated using a bootstrapping approach without replacement by randomly selecting a specific number of daily scores during the week according to the missing day scenario (scores missing for 1, 2, 3, 4, 5, or 6 days). For each missing-day scenario, each participant's simulated weekly item score was calculated as the mean of randomly selected daily scores. The average score across weeks was then calculated for each

participant. Finally, the mean and SD were calculated across participants. To identify the point at which substantial changes in the variability of weekly item scores occurred, the SD for each missing-day scenario was compared with the SD when no days were missing using the Brown–Forsythe test.[20]

Reliability

Internal consistency reliability reflects the extent to which individual items from a scale consisting of multiple items are measuring the same general concept when measured at a single time point. In the present context, Cronbach's alpha[21] was calculated for weekly NTDT-PRO T/W and SoB domain scores with standardisation of variances before and after deletion of individual NTDT-PRO weekly items for the T/W domain score. Values ≥0.70 indicated acceptable internal consistency.[22]

Test–retest reliability is a measure of how consistently an instrument measures a concept at different time points in "stable" participants, and was assessed, at the NTDT-PRO domain level, by calculating the intraclass correlation coefficient (ICC) for weekly domain scores using a 2-way mixed-effects analysis of variance (ANOVA) model with week as a fixed effect.[23] Stable participants were those with PGI-S weekly scores at baseline and week 1 that differed by ≤0.5 points. An ICC of ≥0.70 indicated acceptable test–retest reliability.[24]

Validity

Convergent validity is demonstrated when different measures of the same concept are strongly correlated with each other, while discriminant validity can be inferred when unrelated concepts are weakly correlated. Convergent and discriminant validity was assessed via Spearman's rank correlation coefficients between NTDT-PRO domain scores and other scores (PGI-S score, and domain and summary scores for the SF-36v2® and FACIT-F) from assessments done at the same time point (baseline, week 24, or weeks 13–24). It was hypothesised that NTDT-PRO domain scores would be moderately to strongly related (Spearman's rank correlation coefficient: ≥0.3) to SF-36v2® physical functioning and vitality, FACIT-F physical well-being and FS, and the PGI-S scores, and less

related (Spearman's rank correlation coefficient: <0.3) to SF-36v2® bodily pain, role-emotional, and MCS scores.

Known-groups validity of the NTDT-PRO domains—sensitivity to differentiate among groups of participants known to be clinically different—was assessed by comparing least-squares (LS) mean NTDT-PRO scores between different subgroups of participants, classified based on scores for the PGI-S, the FACIT-F FS, SF-36v2® vitality, and selected FACIT-F items and SF-36v2® items. The domains and items were selected for their theorised relationship to the concepts being measured by the NTDT-PRO T/W and SoB domains. Classifications used to define known groups are shown in online supplementary table S2. Classifications for the PGI-S were defined based on the assumption of a 2-point meaningful difference. For the FACIT-F FS, the cut-off used by the instrument developer to differentiate patients with cancer from the general population was used to classify participants as moderate or mild.[25] A clinically important difference of 3 points, as suggested by instrument developer, was used to define the other categories.[26] The SF-36v2® vitality "normal" category was defined based on a meaningful difference of ±6.7 points from the norm-based mean score of 50, with other categories defined by subsequently adding or subtracting 6.7 from the upper or lower bounds, respectively.[17] For item-based known groups, each verbal response level was taken as a known group. Analysis of covariance (ANCOVA) models were used that included NTDT-PRO domain scores at baseline, week 24, and weeks 13–24 as the dependent variable, and the known-groups measure at the corresponding time point as the independent variable, and that were adjusted for age and geographic region.

Responsiveness

Responsiveness was defined as the sensitivity of the NTDT-PRO to changes in a patient's symptom severity over time. Responsiveness was evaluated by first calculating Spearman rank correlation coefficients for changes from baseline in NTDT-PRO domain scores at week 24 and weeks 13–24 and the changes in haemoglobin level (generally considered as a measure of response) and scores for FACIT-F FS, SF-36v2® vitality, the PGI-S, the PGI-C, and selected FACIT-F and SF-36v2® items. The 5 measures with the strongest correlations at weeks 13–24 with NTDT-PRO domain score

changes were included in a subsequent analysis where ANCOVA models were used to compare LS mean changes in NTDT-PRO domain scores among different response categories. Response categories (table 1) were defined based on reported estimates of clinically meaningful within-patient changes for FACIT-F FS and SF-36v2® vitality domain scores or 1-point differences for individual items. A 1-point difference was also used to define the response categories of the PGI-S. The models included NTDT-PRO domain scores change as the dependent variable and response categories for the given anchor measure as the independent variable, and were adjusted for age and geographic region.

Table 1 Responsiveness at weeks 13–24

	Spearman's rank correlation coefficient (r)[a]		Least-squares mean change (95% CI) at weeks 13–24[b]				
	Week 24	Weeks 13–24	Improvement level 2	Improvement level 1	No change	Worsening	*p* value[c]
NTDT-PRO T/W domain							
Haemoglobin level	–0.38	–0.30	–	–	–	–	–
SF-36v2® vitality	–0.49	–0.46	–	–1.77 (–2.42, –1.12)	–0.40 (–0.80, 0.00)	0.60 (–0.20, 1.39)	<0.001
SF-36v2® item 9e	0.28	0.41	–	–	–	–	–
SF-36v2® item 9g	–0.41	–0.40	–	–	–	–	–
SF-36v2® item 9i	–0.42	–0.43	–	–	–	–	–
FACIT-F FS	–0.52	–0.56	–2.74 (–3.42, –2.06)	–1.68 (–2.44, –0.93)	–0.22 (–0.57, 0.13)	0.42 (–0.16, 1.01)	<0.001
FACIT-F item HI7	–0.41	–0.40	–	–	–	–	–
FACIT-F item HI12	–0.58	–0.60	–3.28 (–4.24, –2.32)	–1.69 (–2.44, –0.95)	–0.51 (–0.88, –0.13)	0.48 (–0.08, 1.03)	<0.001
FACIT-F item An2	–0.43	–0.45	–	–1.84 (–2.46, –1.22)	–0.21 (–0.61, 0.20)	0.00 (–0.68, 0.68)	<0.001

FACIT-F item An5	–0.33	–0.31	–	–	–	–	–
PGI-S	0.83	0.79	–3.26 (–3.75, –2.77)	–1.80 (–2.35, –1.25)	–0.09 (–0.35, 0.18)	0.99 (0.56, 1.42)	<0.001
PGI-C	0.39	0.28	–	–	–	–	–
NTDT-PRO SoB domain							
Haemoglobin level	–0.36	–0.32	–	–	–	–	–
SF-36v2® vitality	–0.40	–0.41	–	–1.28 (–1.91, –0.66)	–0.22 (–0.60, 0.16)	0.52 (–0.24, 1.28)	<0.001
SF-36v2® item 9e	0.30	0.41	–	–	–	–	–
SF-36v2® item 9g	–0.38	–0.36	–	–	–	–	–
SF-36v2® item 9i	–0.30	–0.34	–	–	–	–	–
FACIT-F FS	–0.49	–0.51	–2.21 (–2.88, –1.53)	–1.18 (–1.92, –0.43)	–0.01 (–0.36, 0.33)	0.25 (–0.32, 0.83)	<0.001
FACIT-F item HI7	–0.32	–0.29	–	–	–	–	–
FACIT-F item HI12	–0.45	–0.48	–2.70 (–3.64, –1.76)	–1.08 (–1.81, –0.35)	–0.25 (–0.62, 0.12)	0.33 (–0.22, 0.87)	<0.001
FACIT-F item An2	–0.39	–0.43	–	–1.38 (–1.97, –0.78)	–0.07 (–0.45, 0.32)	0.09 (–0.56, 0.74)	<0.001
FACIT-F item An5	–0.36	–0.31	–	–	–	–	–
PGI-S	0.68	0.69	–2.62 (–3.14, –2.09)	–1.17 (–1.77, –0.58)	0.00 (–0.28, 0.28)	1.01 (0.55, 1.47)	<0.001
PGI-C	0.30	0.28	–	–	–	–	–

[a]Changes from baseline.
[b]Score changes defining response categories (improvement level 2, improvement level 1, no change, worsening): SF-36v2® vitality: N/A, ≥6.7, >–6.7 to <6.7, ≤–6.7; FACIT-F FS: ≥8, 4 to <8, >–4 to <4, ≤–4; FACIT-F item HI12: ≥2, 1 to <2, >–1 to <1, ≤–1; FACIT-F item An2: N/A, ≥1, >–1 to <1, ≤–1; PGI-S: ≤–2, >–2 to –1, >–1 to <1, ≥1. For SF-36v2® vitality and FACIT-F Item An2, no improvement level 2 category was used.
[c]*F*-test comparing T/W and SoB domain scores across response categories (ANCOVA).
ANCOVA, analysis of covariance; CI, confidence interval; FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FS, Fatigue Subscale; N/A, not applicable; PGI-C, Patient Global Impression of Change; PGI-S, Patient Global Impression of Severity; SF-36v2 ®, Short Form Health Survey version 2; SoB, Shortness of Breath; T/W, Tiredness/Weakness.

RESULTS

Participants

The ITT population comprised 145 participants with a mean (SD) age of 39.9 (12.8) years (range: 18 to 71 years) (see online supplementary table S3). Most participants were female (56.6%), White (60.0%), and from North America or Europe (62.1%). A total of 26.9% of participants had a diagnosis of haemoglobin E/β-thalassaemia, and 6.2% had a diagnosis of β-thalassaemia combined with α-thalassaemia. The mean (SD) haemoglobin level at baseline was 8.2 (1.2) g/dL, and most participants had no or only a slight transfusion burden (mean: 0.3 units of red blood cells in the 24 weeks before the first dose of study drug). Most participants (69.0%) had an ECOG performance status of 0, indicating normal functioning.

Quality of completion of the NTDT-PRO

Across all NTDT-PRO items, the percentage of participants with <4 days of missing NTDT-PRO data (i.e., with sufficient data to calculate average weekly item scores) was 98.6% at baseline and 84.4% at week 24 (see online supplementary table S4). Across the first 24 weeks of treatment, at least 87% of participants per week had non-missing NTDT-PRO T/W and SoB scores (see online supplementary figure S1).

PRO score distributions at baseline

Average weekly NTDT-PRO item scores at baseline ranged from 2.4 for item 5-SobNA (shortness of breath not doing physical activity) to 5.0 for item 2-TiredPA (tiredness doing physical activity) (see online supplementary table S5). Baseline average weekly domain scores were 4.1 for T/W and 3.3 for SoB. The weekly average PGI-S score at baseline was 3.7, and average scores for the SF-36v2® scales and component summaries ranged from 42.2 for general health to 51.5 for bodily pain. The average baseline FACIT-F FS score of 36.4 was worse than that in the US general population (43.6).[24] Nonetheless, these data collectively suggested that participants generally had mild to moderate symptoms at study baseline.

Based on skewness and kurtosis values, the distributions of weekly T/W and SoB scores at baseline were generally symmetric but slightly platykurtic, indicating that few participants had extreme values. For T/W, 1.4% of participants had a score of 0 and 1.4% had a score >9; 7.6% of participants had an SoB score of 0 and 0.7% had an SoB score >9 (see online supplementary table S5). For each week up to week 24, <6% of participants had a T/W score of 0, <2% had a T/W score >9, <15% had an SoB score of 0, and <1% had an SoB score >9. This indicates that there were no problematic floor or ceiling effects.

NTDT-PRO item–item and item–domain correlations

Across the 3 assessment time points/time intervals, item 1-TiredNA (tiredness not doing physical activity) was very strongly correlated with item 3-WeakNA (weakness not doing physical activity) (r=0.97 to 0.98), and item 2-TiredPA was very strongly correlated with item 4-WeakPA (weakness doing physical activity) (r=0.98 to 0.99) (table 2). Item 5-SobNA and item 6-SobPA (shortness of breath doing physical activity) were strongly correlated with each other (r=0.74 to 0.81) and moderately to strongly correlated with item 1-TiredNA, item 2-TiredPA, item 3-WeakNA, and item 4-WeakPA (r=0.50 to 0.81).

At the domain level, T/W and SoB scores were strongly correlated with each other (r=0.77 to 0.79). As anticipated, item 1-TiredNA, item 2-TiredPA, item 3-WeakNA, and item 4-WeakPA correlated more strongly with T/W (r=0.88 to 0.95) than with SoB (r=0.67 to 0.77), and item 5-SobNA and item 6-SobPA correlated more strongly with SoB (r=0.89 to 0.97) than with T/W (r=0.64 to 0.78).

Table 2 NTDT-PRO item–item and item–domain correlations

	Spearman's rank correlation coefficient (r)							
	Item 1-TiredNA	**Item 2-TiredPA**	**Item 3-WeakNA**	**Item 4-WeakPA**	**Item 5-SobNA**	**Item-6 SobPA**	**T/W domain**	**SoB domain**
Baseline (N=145)								
Item 1-TiredNA	–	0.77	0.97	0.75	0.75	0.67	0.93	0.75
Item 2-TiredPA	0.77	–	0.73	0.98	0.57	0.77	0.94	0.72

	Spearman's rank correlation coefficient (r)							
	Item 1-TiredNA	**Item 2-TiredPA**	**Item 3-WeakNA**	**Item 4-WeakPA**	**Item 5-SobNA**	**Item-6 SobPA**	**T/W domain**	**SoB domain**
Item 3-WeakNA	0.97	0.73	–	0.74	0.77	0.65	0.91	0.74
Item 4-WeakPA	0.75	0.98	0.74	–	0.58	0.78	0.94	0.73
Item 5-SobNA	0.75	0.57	0.77	0.58	–	0.81	0.70	0.93
Item 6-SobPA	0.67	0.77	0.65	0.78	0.81	–	0.77	0.96
T/W domain	0.93	0.94	0.91	0.94	0.70	0.77	–	0.78
SoB domain	0.75	0.72	0.74	0.73	0.93	0.96	0.78	–
Week 24 (N=110)								
Item 1-TiredNA	–	0.73	0.97	0.71	0.76	0.59	0.89	0.69
Item 2-TiredPA	0.73	–	0.72	0.99	0.54	0.80	0.95	0.75
Item 3-WeakNA	0.97	0.72	–	0.72	0.80	0.62	0.89	0.73
Item 4-WeakPA	0.71	0.99	0.72	–	0.56	0.81	0.95	0.77
Item 5-SobNA	0.76	0.54	0.80	0.56	–	0.75	0.68	0.89
Item 6-SobPA	0.59	0.80	0.62	0.81	0.75	–	0.78	0.97
T/W domain	0.89	0.95	0.89	0.95	0.68	0.78	–	0.79
SoB domain	0.69	0.75	0.73	0.77	0.89	0.97	0.79	–
Weeks 13–24 (N=131)								
Item 1-TiredNA	–	0.71	0.98	0.70	0.73	0.57	0.88	0.67
Item 2-TiredPA	0.71	–	0.71	0.99	0.50	0.79	0.95	0.74
Item 3-WeakNA	0.98	0.71	–	0.72	0.77	0.61	0.89	0.72
Item 4-WeakPA	0.70	0.99	0.72	–	0.52	0.81	0.95	0.76
Item 5-SobNA	0.73	0.50	0.77	0.52	–	0.74	0.64	0.89
Item 6-SobPA	0.57	0.79	0.61	0.81	0.74	–	0.76	0.96
T/W domain	0.88	0.95	0.89	0.95	0.64	0.76	–	0.77
SoB domain	0.67	0.74	0.72	0.76	0.89	0.96	0.77	–

SoB, Shortness of Breath; SobNA, shortness of breath not doing physical activity; SobPA, shortness of breath doing physical activity; TiredNA, tiredness not doing physical activity; TiredPA, tiredness doing physical activity; WeakNA, weakness not doing physical activity; WeakPA, weakness doing physical activity; T/W, Tiredness/Weakness.

Weekly scoring rule

For all NTDT-PRO items, mean scores varied very little between different scenarios where the number of missing days ranged from 0 to 6 (see online supplementary table S6). Moreover, when comparing SD values for the different missing day scenarios using the Browne–Forsythe test, none of the SDs from the missing days were statistically significantly different from the SD when no days were missing. The requirement that scores be available for at least 4 days for a weekly score to be calculated was therefore shown to be reasonable.

Reliability

Internal consistency reliability

Cronbach's alpha for the NTDT-PRO T/W domain was 0.94 to 0.95 across the 3 assessment time points/time intervals (baseline, week 24, weeks 13–24) (see online supplementary table S7), indicating acceptable internal consistency reliability but suggesting possible item redundancy. However, removing individual items from the T/W domain did not increase Cronbach's alpha, indicating that there was no item redundancy. Cronbach's alpha for the NTDT-PRO SoB domain was 0.84 to 0.89, also indicating acceptable internal consistency reliability.

Test–retest reliability

In stable participants (those with a difference in PGI-S weekly scores of ≤0.5 points between baseline and week 1: N=73), ICC was 0.94 for the T/W domain and 0.92 for the SoB domain. These values were comfortably above the prespecified acceptability threshold of 0.70, indicating very good test–retest reliability.

Validity

Convergent and discriminant validity

Hypothesised convergent validity of NTDT-PRO with SF-36v2® physical functioning and vitality, FACIT-F physical well-being, FACIT-F FS, and PGI-S was demonstrated, with all correlation

coefficients exceeding the prespecified threshold of 0.3 in the expected direction (negative for the SF-36v2® and FACIT-F domains and positive for the PGI-S) (table 3). By contrast, with the exception of the weak correlation between SoB and SF-36v2® bodily pain at week 24 (r=–0.29), the hypothesised discriminant validity with SF-36v2® bodily pain, role-emotional, and MCS was not demonstrated.

Table 3 Convergent and discriminant validity

	Spearman's rank correlation coefficient (r)					
	NTDT-PRO T/W domain			NTDT-PRO SoB domain		
	Baseline	Week 24	Weeks 13–24	Baseline	Week 24	Weeks 13–24
SF-36v2®[a]						
Physical functioning	–0.50	–0.35	–0.43	–0.50	–0.35	–0.40
Role-physical	–0.65	–0.44	–0.50	–0.60	–0.40	–0.52
Bodily pain	–0.43	–0.34	–0.41	–0.38	–0.29	–0.37
General health	–0.53	–0.29	–0.34	–0.45	–0.37	–0.36
Vitality	–0.73	–0.61	–0.60	–0.61	–0.56	–0.52
Social functioning	–0.56	–0.34	–0.37	–0.55	–0.32	–0.44
Role-emotional	–0.55	–0.36	–0.43	–0.54	–0.31	–0.47
Mental health	–0.53	–0.38	–0.44	–0.50	–0.37	–0.43
PCS	–0.60	–0.35	–0.44	–0.54	–0.36	–0.43
MCS	–0.62	–0.46	–0.48	–0.58	–0.41	–0.47
FACIT-F[b]						
Physical well-being	–0.69	–0.55	–0.60	–0.60	–0.47	–0.51
Social/family well-being	–0.33	–0.27	–0.23	–0.30	–0.28	–0.22
Emotional well-being	–0.54	–0.35	–0.39	–0.50	–0.40	–0.41
Functional well-being	–0.62	–0.38	–0.42	–0.60	–0.44	–0.39
FACT-G total score	–0.66	–0.46	–0.49	–0.61	–0.47	–0.46
FACIT-F FS	–0.76	–0.58	–0.65	–0.66	–0.55	–0.52
FACIT-F TOI	–0.78	–0.55	–0.64	–0.69	–0.54	–0.54
FACIT-F total score	–0.74	–0.53	–0.58	–0.67	–0.52	–0.51
PGI-S[c]	0.86	0.83	0.80	0.72	0.67	0.65

[a]n=141 at baseline, n=96 at week 24, n=125 at weeks 13–24.
[b]n=144 at baseline, n=96 at week 24, n=126 at weeks 13–24.
[c]n=145 at baseline, n=110 at week 24, n=131 at weeks 13–24.
FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FACT-G, Functional Assessment of Cancer Therapy – General; FS, Fatigue Subscale; MCS, Mental Component Summary; PCS, Physical Component Summary; PGI-S, Patient Global Impression of Severity; SF-36v2 ®, Short Form Health Survey version 2; SoB, Shortness of Breath; TOI, trial outcome index; T/W, Tiredness/Weakness.

Known-groups validity

Known-groups validity was assessed using FACIT-F FS, SF-36v2® vitality, selected FACIT-F and SF-36v2® items, and the PGI-S. The FACIT-F and SF-36v2® items respectively measure similar concepts as the FACIT-F FS and SF-36v2® vitality, but had the advantage of clearly defined rating scales that provided clear cut-off values to differentiate levels of severity. At weeks 13–24 (table 4), as well as at baseline (see online supplementary table S8) and week 24 (see online supplementary table S2), LS mean T/W and SoB scores on the NTDT-PRO were significantly higher (worse) in participants with lower (worse) scores for the FACIT-F FS, FACIT-F items HI12 (feeling weak all over) and An2 (feeling tired), SF-36v2® vitality, and SF-36v2® items 9g (feeling worn out) and 9i (feeling tired), and in participants with higher (worse) scores for SF-36v2® item 9e (having a lot of energy) and the PGI-S. Known-groups validity of the T/W and SoB domains was therefore demonstrated.

Table 4 Known-groups validity at weeks 13–24

		NTDT-PRO T/W domain			NTDT-PRO SoB domain		
	n	**LS mean**	**95% CI**	***p* value[a]**	**LS mean**	**95% CI**	***p* value[a]**
FACIT-F FS				<0.001			<0.001
Very severe (≤37)	43	4.39	3.90, 4.88		3.90	3.35, 4.45	
Severe (>37 to 40)	16	2.91	2.10, 3.73		1.77	0.86, 2.68	
Moderate (>40 to 43)	19	2.81	2.06, 3.55		2.61	1.77, 3.45	
Mild (>43 to 46)	17	1.86	1.05, 2.67		1.92	1.01, 2.83	
Very mild/no symptoms (>46)	31	1.17	0.57, 1.78		0.87	0.19, 1.55	
FACIT-F item HI12[b]				<0.001			<0.001
Very much (0)	5	5.50	4.08, 6.92		3.23	1.60, 4.87	
Quite a bit (1)	16	4.81	4.01, 5.60		4.26	3.34, 5.17	
Somewhat (2)	25	3.70	3.08, 4.33		3.51	2.79, 4.23	
A little bit (3)	53	2.57	2.08, 3.07		2.12	1.55, 2.68	
Not at all (4)	27	1.13	0.48, 1.79		0.84	0.09, 1.59	
FACIT-F item An2[b]				<0.001			<0.001
Very much (0)	8	5.33	4.10, 6.56		3.44	2.07, 4.81	
Quite a bit (1)	12	4.80	3.81, 5.80		4.18	3.08, 5.29	
Somewhat (2)	25	3.38	2.70, 4.07		3.55	2.78, 4.31	
A little bit (3)	64	2.44	1.94, 2.94		1.93	1.37, 2.48	
Not at all (4)	17	1.52	0.66, 2.38		1.20	0.25, 2.16	
SF-36v2® vitality				<0.001			<0.001
Very poor (≤36.6)	20	5.35	4.45, 6.26		4.54	3.54, 5.55	

		NTDT-PRO T/W domain			NTDT-PRO SoB domain		
	n	LS mean	95% CI	*p* value[a]	LS mean	95% CI	*p* value[a]
Poor (>36.6 to 43.3)	19	4.51	3.54, 5.48		3.83	2.76, 4.89	
Normal (>43.3 to 56.7)	64	3.05	2.55, 3.55		2.82	2.27, 3.37	
Better (>56.7 to 63.4)	25	1.86	1.29, 2.44		1.34	0.70, 1.98	
Much better (>63.4)	13	2.45	1.17, 3.73		2.14	0.72, 3.55	
SF-36v2® item 9e[c]				<0.001			<0.001
All of the time (1)	8	2.50	1.29, 3.71		1.69	0.32, 3.06	
Most of the time (2)	44	1.82	1.27, 2.36		1.69	1.07, 2.31	
Some of the time (3)	45	3.18	2.66, 3.70		2.65	2.06, 3.24	
A little of the time (4)	22	4.62	3.87, 5.37		4.43	3.58, 5.28	
None of the time (5)	6	5.64	4.28, 7.01		3.69	2.13, 5.24	
SF-36v2® item 9g[c]				<0.001			<0.001
All of the time (1)	4	5.92	4.30, 7.54		4.37	2.56, 6.19	
Most of the time (2)	11	5.30	4.31, 6.29		4.43	3.32, 5.53	
Some of the time (3)	34	3.49	2.93, 4.06		3.17	2.54, 3.80	
A little of the time (4)	49	2.67	2.16, 3.19		2.45	1.87, 3.03	
None of the time (5)	27	1.43	0.77, 2.09		0.83	0.09, 1.56	
SF-36v2® item 9i[c]				<0.001			<0.001
All of the time (1)	7	5.37	4.01, 6.73		4.01	2.51, 5.51	
Most of the time (2)	25	4.32	3.60, 5.05		3.88	3.08, 4.68	
Some of the time (3)	38	2.88	2.29, 3.47		2.55	1.90, 3.20	
A little of the time (4)	49	2.17	1.62, 2.73		1.72	1.11, 2.34	
None of the time (5)	6	2.21	0.76, 3.67		2.14	0.53, 3.74	
PGI-S				<0.001			<0.001
0 to 2 (no symptoms)	45	1.37	0.94, 1.79		1.10	0.57, 1.62	
>2 to 4 (mild)	36	2.93	2.47, 3.40		2.68	2.10, 3.26	
>4 to 6 (moderate)	34	4.48	3.99, 4.98		3.95	3.32, 4.57	
>6 to 8 (severe)	11	4.94	4.16, 5.73		4.18	3.20, 5.17	
>8 (very severe)	5	6.82	5.65, 7.98		5.91	4.45, 7.38	

[a]*F*-test comparing T/W and SoB domain scores across subgroups (ANCOVA).
[b]"Please select one answer […] to indicate your response as it applies to the past 7 days": item HI12, "I feel weak all over"; item An2, "I feel tired".
[c]"How much of the time during the past week did you…": item 9e, "…have a lot of energy?"; item 9g, "…feel worn out?"; item 9i, "…feel tired?"
ANCOVA, analysis of covariance; CI, confidence interval; FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FS, Fatigue Subscale; LS, least-squares; PGI-S, Patient Global Impression of Severity; SF-36v2®, Short Form Health Survey version 2; SoB, Shortness of Breath; T/W, Tiredness/Weakness.

Responsiveness

Considering changes from baseline to week 24 and weeks 13–24, NTDT-PRO T/W and SoB domain scores were moderately correlated with changes in haemoglobin level (–0.30 to –0.38) and weakly to

moderately correlated with the PGI-C (0.28 to 0.39) (table 1). The strongest correlations for the T/W and SoB domain score changes were with changes on SF-36v2® vitality (–0.40 to –0.49), the FACIT-F FS (–0.49 to –0.56), FACIT-F items HI12 (feeling weak all over, –0.45 to –0.60) and An2 (feeling tired, –0.39 to –0.45), and the PGI-S (0. 68 to 0.83). In a responsiveness analysis using these 5 measures as anchors, decreases (improvements) in LS mean T/W and SoB scores were significantly higher in participants with greater improvements in scores on the anchors. The T/W and SoB domains were therefore shown to be responsive to changes in symptom severity (table 1).

DISCUSSION

Broadly, the NTDT-PRO demonstrated sufficient psychometric performance to defend its use as a measure of treatment outcome in clinical research among patients with NTDT. Distributional properties were good, as illustrated by the lack of floor and ceiling effects. High ICC values in patients assessed as stable based on PGI-S scores at baseline and week 1 indicated good test–retest reliability, while similarly high Cronbach's alpha coefficients at baseline, week 24, and weeks 13–24 indicated good internal consistency reliability. Correlation analyses confirmed the hypothesised direction and strength of relationship of both NTDT-PRO domains with other PRO measures, although the hypothesised discriminant validity with SF-36v2® bodily pain, role-emotional, and MCS was not demonstrated. However, as weakness, tiredness, and shortness of breath are broad concepts, it was not wholly surprising that NTDT-PRO T/W and SoB domain scores were correlated with these SF-36v2® scores. Finally, known-groups validity and responsiveness were demonstrated based on the PGI-S and selected FACIT-F and SF-36v2® items.

These findings build on an earlier preliminary psychometric analysis using data from 48 adults with NTDT who participated in a multicentre observational study, which demonstrated that the NTDT-PRO had high internal consistency reliability and test–retest reliability.[15] That earlier study was unable to adequately evaluate sensitivity to change, however, due to its non-interventional study design. This resulted in very few participants experiencing improvement in symptoms, as assessed by the PGI-C. In the present analysis, using data from the first 24 weeks of treatment in the BEYOND trial, the relationship among changes in NTDT-PRO scores relative to changes observed in multiple

other measures of similar and distinct concepts at week 24 and weeks 13–24 were as we hypothesised, and are supportive of the tool's ability to detect change.

Although the NTDT-PRO T/W and SoB domains were shown to be responsive to changes over time on all the anchors examined in the responsiveness analysis, changes in the PGI-C had the weakest correlation (0.28) with change in T/W domain score at weeks 13–24 among the included anchors. The weaker correlation between the NTDT-PRO domain score changes and the PGI-C as compared to other potential anchors may be due to an issue with recall: it may have been difficult for patients to rate how much their overall thalassaemia symptoms—which can be many—had changed in the 24 weeks since the beginning of the study.[27,28]

Limitations of the present study include the modest sample size for typical psychometric evaluations, although it was adequate for assessment of the trial endpoints. NTDT is a rare disease, which makes recruitment challenging. Moreover, cut-off values defining different levels of improvement are not yet well established for some of the anchors included in the responsiveness analysis (PGI-S, FACIT-F FS, and SF-36v2® vitality), so the cut-off values used in the responsiveness analysis were necessarily based on certain assumptions. However, given that score changes for these PRO measures were moderately to strongly correlated with score changes for the NTDT-PRO domains, modifying the cut-off values used to define different levels of improvement would likely yield very similar findings. Strengths of this study include use of well-validated PRO instruments, including the SF-36v2® and FACIT-F. Additionally, data for this analysis were from a phase 2 interventional study with participants from multiple geographic regions and spanning a range of NTDT symptom severities based on baseline T/W and SoB domain scores. This confirms the validity of the NTDT-PRO over a broad population. The use of data from an interventional study also allowed for changes in symptom severity to be observed, a necessity for validating the sensitivity to change of the NTDT-PRO domains.

In conclusion, the NTDT-PRO demonstrated adequate reliability, validity, and responsiveness when used to assess tiredness/weakness and shortness of breath in patients with NTDT. As a fully validated PRO instrument, it can be used to confidently assess the efficacy of treatments targeting anaemia in

clinical studies for NTDT. Future analyses will focus on the NTDT-PRO score interpretability by identifying meaningful change thresholds and symptomatic thresholds for the T/W and SoB domains.

Acknowledgments The authors received medical writing support in the preparation of this manuscript from Stephen Gilliver, PhD, of Evidera, and editorial support from Patricia Fonseca, PhD, of Excerpta Medica, funded by Bristol Myers Squibb.

Data availability statement

The data that supports the findings of this study are available in the supplementary material of this article and the data that support the findings of this study are available from the corresponding author.

Contributors

SG, CP and AS designed and conceptualised the study. SG and CP analysed the data. All authors critically interpreted the data and revised the article. AT is responsible for the overall content as the corresponding author. All authors contributed to critically editing and approving the final manuscript.

Funding

This study was funded by Bristol Myers Squibb. Award/grant number: not applicable.

Competing interests

A.T.T.: consulting fees from Agios Pharmaceuticals; research funding and consulting fees from Celgene/Bristol Myers Squibb, Ionis Pharmaceuticals, Novartis Pharmaceuticals, and Vifor Pharma. K.M.M.: consulting fees from Agios Pharmaceuticals, Celgene/Bristol Myers Squibb, CRISPR Therapeutics, Novartis, Pharmacosmos, and Vifor Pharma. V.V.: research funding from Bristol Myers Squibb. A.K.: advisory board fees and consulting fees from Agios Pharmaceuticals, Celgene/Bristol Myers Squibb, Chiesi Farmaceutici, CRISPR Therapeutics/Vertex Pharmaceuticals, Ionis Pharmaceuticals, Novartis, and Vifor Pharma; research support from Celgene/Bristol Myers Squibb and Novartis. J.L.-B., A.Y., J.K.S., and L.M.B.: employment by and stock/equity holder of Bristol

Myers Squibb. S.G.: employment by Evidera; consultancy fees from Bristol Myers Squibb, Gilead, and Janssen. C.P.: employment by Evidera. A.L.S.: employment by Adelphi Values. D.M.: employment by Bristol Myers Squibb. M.D.C.: advisory board fees from Celgene/Bristol Myers Squibb, CRISPR Therapeutics, Ionis Pharmaceuticals, Novartis, Novo Nordisk, Sanofi Genzyme, and Vifor Pharma.

Ethics approval

The BEYOND trial received institutional review board/ethics committee approval (sites 101 and 102, A. Kattamis and E. Voskaridou: 112/17; site 201, MD Cappellini: CE150176; site 202: GL Forni: CE150176 and CE150124; site 203, S Perrotta: CE150176 and CE150110; site 204, AG Piga: CE150176 and CE150089; site 206, A Filosa: CE150176 and CE150040; site 301, AT Taher: NA and BIO-2017-0338; site 401: V Viprakasit: 689/2560(EC4); site 501, TD Coates: CHLA-17-00444; site 503, AA Thompson: IRB 2018-1580; and site 601, JB Porter: 17/EM/0438) and was conducted in accordance with International Council for Harmonisation Good Clinical Practice and the Declaration of Helsinki.

Patient consent for publication

Not required for this analysis.

Clinical trial registration

ClinicalTrials.gov Identifier: NCT03342404 (BEYOND)

REFERENCES

1 Taher AT, Musallam KM, Cappellini MD. Beta-thalassemias. *N Engl J Med* 2021;384:727–43.

2 Musallam KM, Rivella S, Vichinsky E, *et al*. Non-transfusion-dependent thalassemias. *Haematologica.* 2013;98(6):833-844.

3 Musallam KM, Cappellini MD, Viprakasit V, *et al*. Revisiting the non-transfusion-dependent (NTDT) vs. transfusion-dependent (TDT) thalassemia classification 10 years later. *Am J Hematol* 2021;96:E54–6.

4 Taher AT, Musallam KM, El-Beshlawy A, *et al*. Age-related complications in treatment-naive patients with thalassaemia intermedia. *Br J Haematol* 2010;150:486–9.

5 Musallam KM, Vitrano A, Meloni A, et al. Survival and causes of death in 2,033 patients with non-transfusion-dependent β-thalassemia. *Haematologica.* 2021;106(9):2489-2492.

6 Musallam KM, Vitrano A, Meloni A, *et al*. Risk of mortality from anemia and iron overload in nontransfusion-dependent β-thalassemia. *Am J Hematol* 2022;97:E78–80.

7 Arian M, Mirmohammadkhani M, Ghorbani R, *et al*. Health-related quality of life (HRQoL) in beta-thalassemia major (β-TM) patients assessed by 36-item short form health survey (SF-36): a meta-analysis. *Qual Life Res* 2019;28:321–34.

8 Telfer P, Constantinidou G, Andreou P, *et al*. Quality of life in thalassemia. *Ann N Y Acad Sci* 2005;1054:273–82.

9 Lyrakos GN, Vini D, Aslani H, *et al*. Psychometric properties of the Specific Thalassemia Quality of Life Instrument for adults. *Patient Prefer Adherence* 2012;6:477–97.

10 Klaassen RJ, Barrowman N, Merelles-Pulcini M, *et al*. Validation and reliability of a disease-specific quality of life measure (the TranQol) in adults and children with thalassaemia major. *Br J Haematol* 2014;164:431–7.

11 Musallam KM, Khoury B, Abi-Habib R, *et al*. Health-related quality of life in adults with transfusion-independent thalassaemia intermedia compared to regularly transfused thalassaemia major: new insights. *Eur J Haematol* 2011;87:73–9.

12 Khoury B, Musallam KM, Abi-Habib R, *et al*. Prevalence of depression and anxiety in adult patients with β-thalassemia major and intermedia. *Int J Psychiatry Med* 2012;44:291–303.

13 Taher A, Viprakasit V, Cappellini MD, *et al*. Development of a patient-reported outcomes symptom measure for patients with nontransfusion-dependent thalassemia (NTDT-PRO©). *Am J Hematol* 2019;94:171–6.

14 FDA. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. 2009; https://www.fda.gov/media/77832/download (accessed 16 June 2022).

15 Taher A, Cappellini MD, Viprakasit V, *et al*. Validation of a patient-reported outcomes symptom measure for patients with nontransfusion-dependent thalassemia (NTDT-PRO©). *Am J Hematol* 2019;94:177–83.

16. Taher AT, Cappellini MD, Kattamis A, et al. The BEYOND Study: Results of a Phase 2, Double-Blind, Randomized, Placebo-Controlled Multicenter Study of Luspatercept in Adult Patients With Non-Transfusion Dependent Beta-Thalassemia. *HemaSphere* 2021;5(S2):1–2.

17. Maruish ME. *User's Manual for the SF-36v2 Health Survey.* 3rd ed. Lincoln, RI: QualityMetric; 2011.

18. FACIT Group. FACIT-F scoring guidelines (version 4). 2021; www.facit.org/measures-scoring-downloads/facit-f-scoring-downloads (accessed 28 September 2021).

19. Hinkle DE, Wiersma W, Jurs SG. Applied Statistics for the Behavioral Sciences. 5th ed. Boston, MA: Houghton Mifflin 2003.

20. Brown MB, Forsythe AB. Robust tests for the equality of variances. *J Am Stat Assoc* 1974;69:364–7.

21. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297–334.

22. Aaronson N, Alonso J, Burnam A, *et al*. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res* 2002;11:193–205.

23. Qin S, Nelson L, McLeod, L, *et al*. Assessing test-retest reliability of patient-reported outcome measures using intraclass correlation coefficients: recommendations for selecting and documenting the analytical formula. *Qual Life Res* 2019:28:1029–33.
24. Cappelleri JC, Zou KH, Bushmakin AG, *et al*. *Patient-Reported Outcomes: Measurement, Implementation and Interpretation.* Boca Raton, FL: CRC Press/Taylor & Francis 2014.
25. Cella D, Lai JS, Chang CH, *et al*. Fatigue in cancer patients compared with fatigue in the general United States population. *Cancer* 2002;94:528–38.
26. Cella D, Eton DT, Lai JS, *et al*. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. *J Pain Symptom Manage* 2002;24:547–61.
27. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J Clin Epidemiol* 1997;50:869–79.
28. Nixon A, Doll H, Kerr C, *et al*. Interpreting change from patient reported outcome (PRO) endpoints: patient global ratings of concept versus patient global ratings of change, a case study among osteoporosis patients. *Health Qual Life Outcomes* 2016;14:25.

SUPPLEMENTARY MATERIALS

TABLE S1 ALGORITHM FOR MAPPING PRO ASSESSMENTS TO NOMINAL WEEKS

	Nominal week	NTDT-PRO	FACIT-F/SF-36v2®
Baseline	0	Days –7 to –1	Day of dosing of the first dose of study drug (screening if missing)
Weeks 1–12	1	Days 1 to 7	None
	2	Days 8 to 14	None
	3	Days 15 to 21	None
	4	Days 22 to 28	None
	5	Days 29 to 35	None
	6	Days 36 to 42	Days 22 to 63
	7	Days 43 to 49	None
	8	Days 50 to 56	None
	9	Days 57 to 63	None
	10	Days 64 to 70	None
	11	Days 71 to 77	None
	12	Days 78 to 84	Days 64 to Day 105
Weeks 13–24	13	Days 85 to 91	None
	14	Days 92 to 98	None
	15	Days 99 to 105	None
	16	Days 106 to 112	None
	17	Days 113 to 119	None
	18	Days 120 to 126	Days 106 to 147
	19	Days 127 to 133	None
	20	Days 134 to 140	None
	21	Days 141 to 147	None
	22	Days 148 to 154	None
	23	Days 155 to 161	None
	24	Days 162 to 168	Days 148 to 189

FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; PRO, patient-reported outcomes; SF-36v2®, Short Form Health Survey version 2.

Table S2 Known-groups validity at week 24

	n	NTDT-PRO T/W domain			NTDT-PRO SoB domain		
		LS mean	95% CI	*p* value[a]	LS mean	95% CI	*p* value[a]
FACIT-F FS				<0.001			<0.001
Very severe (≤37)	62	4.04	3.39, 4.69		3.67	2.99, 4.36	
Severe (>37 to 40)	16	2.63	1.61, 3.65		2.14	1.06, 3.22	
Moderate (>40 to 43)	18	2.52	1.59, 3.45		2.50	1.52, 3.48	
Mild (>43 to 46)	17	2.31	1.40, 3.23		2.01	1.04, 2.98	
Very mild/no symptoms (>46)	31	1.05	0.27, 1.82		0.62	–0.21, 1.44	
FACIT-F item HI12[b]				<0.001			<0.001
Very much (0)	3	6.57	4.68, 8.46		4.93	2.79, 7.07	
Quite a bit (1)	10	4.44	3.39, 5.49		3.85	2.67, 5.04	
Somewhat (2)	16	3.29	2.45, 4.12		3.39	2.44, 4.33	
A little bit (3)	40	2.77	2.20, 3.34		2.36	1.72, 3.00	
Not at all (4)	27	1.23	0.55, 1.92		0.93	0.16, 1.71	
FACIT-F item An2[b]				<0.001			0.002
Very much (0)	3	6.62	4.57, 8.68		4.92	2.68, 7.17	
Quite a bit (1)	11	4.08	3.01, 5.16		3.41	2.23, 4.58	
Somewhat (2)	15	3.36	2.45, 4.27		3.59	2.59, 4.58	
A little bit (3)	48	2.34	1.76, 2.93		1.96	1.32, 2.60	
Not at all (4)	19	1.78	0.91, 2.65		1.31	0.36, 2.26	
SF-36v2® vitality				<0.001			<0.001
Very poor (≤36.6)	7	5.37	4.07, 6.67		4.53	3.10, 5.96	
Poor (>36.6 to 43.3)	11	4.45	3.41, 5.49		4.04	2.90, 5.18	
Normal (>43.3 to 56.7)	41	2.98	2.40, 3.56		2.79	2.15, 3.43	
Better (>56.7 to 63.4)	29	1.72	1.05, 2.39		1.25	0.51, 1.98	
Much better (>63.4)	8	1.56	0.31, 2.80		1.48	0.11, 2.84	
SF-36v2® item 9e[c]				<0.001			0.001
All of the time (1)	3	3.13	1.10, 5.17		1.55	–0.72, 3.82	
Most of the time (2)	40	1.79	1.20, 2.39		1.58	0.92, 2.25	
Some of the time (3)	30	2.99	2.34, 3.64		2.76	2.03, 3.48	
A little of the time (4)	15	4.06	3.12, 5.00		3.51	2.47, 4.56	
None of the time (5)	8	5.13	3.88, 6.39		4.44	3.04, 5.85	
SF-36v2® item 9g[c]				<0.001			<0.001
All of the time (1)	5	5.67	4.24, 7.09		4.67	3.11, 6.24	
Most of the time (2)	4	5.03	3.35, 6.71		4.58	2.74, 6.43	
Some of the time (3)	18	3.79	3.01, 4.58		3.57	2.71, 4.43	
A little of the time (4)	44	2.62	2.07, 3.16		2.37	1.77, 2.97	
None of the time (5)	25	1.20	0.51, 1.90		0.78	0.02, 1.54	
SF-36v2® item 9i[c]				<0.001			<0.001
All of the time (1)	3	6.20	4.23, 8.17		6.47	4.30, 8.64	

	n	NTDT-PRO T/W domain			NTDT-PRO SoB domain		
		LS mean	95% CI	*p* value[a]	LS mean	95% CI	*p* value[a]
Most of the time (2)	17	4.36	3.53, 5.19		3.56	2.64, 4.47	
Some of the time (3)	25	2.77	2.03, 3.50		2.53	1.72, 3.34	
A little of the time (4)	44	1.99	1.42, 2.56		1.76	1.14, 2.39	
None of the time (5)	7	1.58	0.25, 2.91		1.49	0.02, 2.96	
PGI-S				<0.001			<0.001
0 to 2 (no symptoms)	43	1.13	0.72, 1.54		0.93	0.37, 1.48	
>2 to 4 (mild)	33	3.43	2.97, 3.89		3.32	2.69, 3.94	
>4 to 6 (moderate)	21	4.31	3.70, 4.91		3.63	2.82, 4.44	
>6 to 8 (severe)	11	5.60	4.85, 6.34		4.99	3.99, 6.00	
>8 (very severe)	2	6.81	5.07, 8.55		4.34	1.99, 6.69	

[a]*F*-test comparing T/W and SoB domain scores across subgroups (ANCOVA).
[b]"Please select one answer […] to indicate your response as it applies to the past 7 days": item HI12, "I feel weak all over"; item An2, "I feel tired".
[c]"How much of the time during the past week did you…": item 9e, "…have a lot of energy?"; item 9g, "…feel worn out?"; item 9i, "…feel tired?"
ANCOVA, analysis of covariance; CI, confidence interval; FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FS, Fatigue Subscale; LS, least-squares; PGI-S, Patient Global Impression of Severity; SF-36v2®, Short Form Health Survey version 2; SoB, Shortness of Breath; T/W, Tiredness/Weakness.

Table S3 Demographics and baseline clinical characteristics

	N=145
Age (years)	
Mean (SD)	39.9 (12.8)
Median (range)	40 (18 to 71)
Female, n (%)	82 (56.6)
Race, n (%)	
Asian	44 (30.3)
White	87 (60.0)
Other	14 (9.7)
Ethnicity, n (%)	
Hispanic or Latino	3 (2.1)
Not Hispanic or Latino	142 (97.9)
Body mass index (kg/m^2), n (%)	
<20	53 (36.6)
20 to <25	66 (45.5)
25 to <30	21 (14.5)
≥30	5 (3.5)
Geographic region, n (%)	
North America and Europe	90 (62.1)
Middle East	17 (11.7)
Asia Pacific	38 (26.2)
β-thalassaemia diagnosis, n (%)	
β-thalassaemia	97 (66.9)
Haemoglobin E/β-thalassaemia	39 (26.9)
β-thalassaemia plus α-thalassaemia	9 (6.2)
Baseline haemoglobin level (g/dL)	
Mean (SD)	8.2 (1.2)
Median (range)	8.2 (7.3 to 9.2)
Categories of baseline haemoglobin level, n (%)	
≥8.5 g/dL	60 (41.4)
<8.5 g/dL	85 (58.6)
Baseline transfusion burden (units of red blood cells in the 24 weeks before the first dose of study drug)	
Mean (SD)	0.3 (0.9)
Median (range)	0 (0 to 6)
6-minute walk test, n (%)	
≤450 m	82 (56.6)
>450 m	63 (43.4)
Left ventricular ejection fraction (%)	
Mean (SD)	65.6 (5.5)
Median (range)	65.0 (55.4 to 79.0)
Tricuspid valve regurgitation velocity, n (%)	
≤2.8 m/s (low probability of pulmonary hypertension)	111 (76.6)
>3.4 m/s (high probability of pulmonary hypertension)	1 (0.7)
ECOG performance status, n (%)	
0	100 (69.0)
1	45 (31.0)

ECOG, Eastern Cooperative Oncology Group; SD, standard deviation.

Table S4 Completeness of NTDT-PRO item entry at baseline and week 24

Number of days with missing NTDT-PRO data[a]	n (%)	
	Baseline (N=145)	Week 24 (N=128)
0	56 (38.6)	51 (39.8)
1	44 (30.3)	31 (24.2)
2	24 (16.6)	20 (15.6)
3	19 (13.1)	6 (4.7)
4	1 (0.7)	10 (7.8)
5	1 (0.7)	7 (5.5)
6	0	3 (2.3)
7	0	0

[a]There was no item-level missing data (participants either completed all 6 NTDT-PRO items or none of them).

Table S5 Baseline PRO score distributions

	Mean (SD)	Median (Q1, Q3)	Range	Skewness	Kurtosis	Floor effect (%)[a]	Ceiling effect (%)[b]
NTDT-PRO							
Item 1-TiredNA	3.2 (2.2)	3.0 (1.5, 4.8)	0.0 to 9.0	0.2	–0.6	11.7	0.0
Item 2-TiredPA	5.0 (2.5)	5.2 (3.4, 7.0)	0.0 to 10.0	–0.3	–0.7	1.4	2.1
Item 3-WeakNA	3.1 (2.2)	3.0 (1.3, 4.8)	0.0 to 9.3	0.3	–0.5	11.7	0.7
Item 4-WeakPA	4.9 (2.6)	5.0 (3.0, 7.0)	0.0 to 10.0	–0.2	–0.8	2.8	2.1
Item 5-SobNA	2.4 (2.1)	2.2 (0.3, 4.0)	0.0 to 8.9	0.7	–0.2	20.7	0.0
Item 6-SobPA	4.2 (2.7)	4.4 (2.0, 6.4)	0.0 to 10.0	0.1	–1.0	7.6	2.8
T/W domain (items 1 to 4)	4.1 (2.2)	4.3 (2.5, 5.7)	0.0 to 9.5	0.0	–0.6	1.4	1.4
SoB domain (items 5 and 6)	3.3 (2.3)	3.4 (1.2, 5.1)	0.0 to 9.4	0.2	–0.8	7.6	0.7
PGI-S	3.7 (2.4)	3.8 (1.8, 5.4)	0.0 to 9.5	0.1	–0.8		
SF-36v2®							
Physical functioning	47.7 (7.7)	48.0 (44.2, 53.7)	23.1 to 57.5	–0.8	0.2	–	–
Role-physical	47.6 (7.8)	48.2 (41.4, 54.9)	25.7 to 57.2	–0.4	–0.7	–	–
Bodily pain	51.5 (9.2)	51.5 (42.6, 62.0)	30.6 to 62.0	–0.3	–1.1	–	–
General health	42.2 (10.2)	41.3 (34.2, 50.8)	19.0 to 66.5	0.1	–0.6	–	–
Vitality	49.2 (10.6)	49.6 (40.7, 58.5)	25.9 to 70.4	–0.3	–0.9	–	–
Social functioning	46.7 (9.3)	47.3 (37.3, 57.3)	22.3 to 57.3	–0.5	–0.8	–	–
Role-emotional	46.6 (8.8)	49.2 (38.8, 52.7)	17.9 to 56.2	–0.7	–0.4	–	–
Mental health	47.2 (9.6)	48.3 (40.4, 56.1)	24.7 to 64.0	–0.5	–0.6	–	–
PCS	48.0 (7.1)	48.8 (43.1, 53.3)	28.4 to 63.6	–0.4	–0.1	–	–
MCS	46.9 (9.2)	47.7 (40.6, 53.9)	23.3 to 63.1	–0.5	–0.4	–	–
FACIT-F							
Physical well-being	22.9 (3.9)	24.0 (20.0, 26.0)	11.0 to 28.0	–0.8	0.0	–	–
Social/family well-being	19.4 (5.3)	20.0 (16.3, 23.0)	4.7 to 28.0	–0.4	–0.5	–	–
Emotional well-being	18.2 (3.5)	19.0 (16.0, 21.0)	8.0 to 24.0	–0.6	–0.4	–	–
Functional well-being	18.0 (5.4)	18.0 (14.0, 22.0)	3.0 to 28.0	0.0	–0.6	–	–
FACT-G total score	78.4 (14.6)	80.0 (67.0, 90.3)	42.0 to 105.8	–0.1	–0.7	–	–
FACIT-F FS	36.4 (9.9)	39.0 (29.0, 44.5)	1.0 to 51.0	–0.7	0.0	–	–
FACIT-F TOI	77.2 (17.2)	81.0 (64.0, 91.0)	29.0 to 105.0	–0.4	–0.7	–	–
FACIT-F total score	114.8 (22.8)	118.5 (100.0, 133.2)	62.0 to 155.8	–0.3	–0.7	–	–

[a]Score of 0.

[b]Score of >9.

FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FACT-G, Functional Assessment of Cancer Therapy – General; FS, Fatigue Subscale; MCS, Mental Component Summary; PCS, Physical Component Summary; PGI-S, Patient Global Impression of Severity; PRO, patient-reported outcomes; Q1, first quartile; Q3, third quartile; SD, standard deviation; SF-36v2®, Short Form Health Survey version 2; SoB, Shortness of Breath; SobNA, shortness of breath not doing physical activity; SobPA, shortness of breath doing physical activity; TiredNA, tiredness not doing physical activity; TiredPA, tiredness doing physical activity; TOI, trial outcome index; T/W, Tiredness/Weakness; WeakNA, weakness not doing physical activity; WeakPA, weakness doing physical activity.

Table S6 Variability of weekly NTDT-PRO item scores across missing day scenarios

		Number of missing days						
		0	**1**	**2**	**3**	**4**	**5**	**6**
Item 1-TiredNA	Mean	2.36	2.36	2.37	2.39	2.31	2.33	2.30
	SD	1.913	1.913	1.917	1.908	1.930	1.931	1.947
	p value[a]	–	0.971	0.949	0.971	0.962	0.869	0.962
Item 2-TiredPA	Mean	4.44	4.44	4.44	4.42	4.46	4.44	4.45
	SD	2.315	2.319	2.308	2.316	2.328	2.352	2.338
	p value[a]	–	1.000	0.953	0.970	0.978	0.827	0.873
Item 3-WeakNA	Mean	2.60	2.60	2.61	2.61	2.59	2.58	2.60
	SD	1.879	1.872	1.872	1.877	1.895	1.917	1.961
	p value[a]	–	0.941	0.930	0.955	0.888	0.786	0.576
Item 4-WeakPA	Mean	4.42	4.42	4.42	4.40	4.44	4.43	4.44
	SD	2.378	2.381	2.392	2.396	2.365	2.369	2.416
	p value[a]	–	0.997	0.973	0.892	0.871	0.965	0.764
Item 5-SobNA	Mean	2.02	2.02	2.01	2.03	2.01	2.05	2.05
	SD	1.894	1.892	1.884	1.911	1.884	1.939	1.928
	p value[a]	–	0.997	0.940	0.911	0.945	0.772	0.788
Item 6-SobPA	Mean	3.76	3.77	3.75	3.76	3.76	3.79	3.74
	SD	2.547	2.546	2.546	2.555	2.548	2.566	2.596
	p value[a]	–	0.982	0.970	0.958	0.993	0.859	0.849

The mean and SD were calculated by first calculating the average score across all weeks for each participant and then calculating the mean and SD across participants.

[a]Brown–Forsythe test comparing SD values for individual missing day scenarios with the SD when 0 days were missing.

SD, standard deviation; SobNA, shortness of breath not doing physical activity; SobPA, shortness of breath doing physical activity; TiredNA, tiredness not doing physical activity; TiredPA, tiredness doing physical activity; WeakNA, weakness not doing physical activity; WeakPA, weakness doing physical activity.

Table S7 NTDT-PRO internal consistency reliability

	Domain	Cronbach's alpha	Deleted item[a]	Cronbach's alpha
Baseline	T/W	0.95		
			Item 1-TiredNA	0.93
			Item 2-TiredPA	0.94
			Item 3-WeakNA	0.94
			Item 4-WeakPA	0.94
	SoB	0.89		
Week 24	T/W	0.94		
			Item 1-TiredNA	0.92
			Item 2-TiredPA	0.92
			Item 3-WeakNA	0.92
			Item 4-WeakPA	0.92
	SoB	0.85		
Weeks 13–24	T/W	0.95		
			Item 1-TiredNA	0.93
			Item 2-TiredPA	0.93
			Item 3-WeakNA	0.93
			Item 4-WeakPA	0.93
	SoB	0.84		

[a]The effect of removing individual items could not be evaluated for the SoB domain, because it consists of only 2 items.

SoB, Shortness of Breath; TiredNA, tiredness not doing physical activity; TiredPA, tiredness doing physical activity; WeakNA, weakness not doing physical activity; WeakPA, weakness doing physical activity; T/W, Tiredness/Weakness.

Table S8 Known-groups validity at baseline

		NTDT-PRO T/W domain			NTDT-PRO SoB domain		
	n	**LS mean**	**95% CI**	***p* value[a]**	**LS mean**	**95% CI**	***p* value[a]**
FACIT-F FS				<0.001			<0.001
Very severe (≤37)	62	5.27	4.84, 5.71		4.35	3.79, 4.91	
Severe (>37 to 40)	16	3.06	2.33, 3.80		3.30	2.36, 4.24	
Moderate (>40 to 43)	18	3.16	2.45, 3.86		2.84	1.93, 3.75	
Mild (>43 to 46)	17	2.94	2.21, 3.68		1.74	0.79, 2.68	
Very mild/no symptoms (>46)	31	1.59	1.05, 2.13		1.13	0.44, 1.83	
FACIT-F item HI12[b]				<0.001			<0.001
Very much (0)	3	7.11	5.47, 8.75		6.23	4.10, 8.36	
Quite a bit (1)	25	5.76	5.16, 6.35		4.80	4.03, 5.57	
Somewhat (2)	24	4.69	4.04, 5.34		4.06	3.22, 4.90	
A little bit (3)	54	3.58	3.18, 3.99		3.08	2.55, 3.60	
Not at all (4)	38	1.71	1.23, 2.18		1.15	0.54, 1.77	
FACIT-F item An2[b]				<0.001			<0.001
Very much (0)	3	7.87	6.21, 9.54		8.02	5.91, 10.13	
Quite a bit (1)	25	5.87	5.26, 6.48		4.89	4.11, 5.66	
Somewhat (2)	37	4.31	3.79, 4.83		3.90	3.24, 4.56	
A little bit (3)	59	3.08	2.68, 3.48		2.31	1.80, 2.82	
Not at all (4)	20	1.43	0.79, 2.08		1.26	0.44, 2.08	
SF-36v2® vitality				<0.001			<0.001
Very poor (≤36.6)	20	6.14	5.43, 6.84		5.57	4.66, 6.48	
Poor (>36.6 to 43.3)	19	5.42	4.70, 6.15		4.11	3.17, 5.05	
Normal (>43.3 to 56.7)	64	3.73	3.32, 4.13		3.15	2.63, 3.68	
Better (>56.7 to 63.4)	25	2.09	1.48, 2.69		1.73	0.95, 2.51	
Much better (>63.4)	13	1.71	0.90, 2.52		1.12	0.07, 2.17	
SF-36v2® item 9e[c]				<0.001			<0.001
All of the time (1)	11	2.09	1.14, 3.04		1.17	–0.02, 2.37	
Most of the time (2)	33	2.21	1.64, 2.77		1.95	1.24, 2.65	
Some of the time (3)	46	3.79	3.27, 4.31		3.24	2.59, 3.89	
A little of the time (4)	37	5.12	4.52, 5.73		4.18	3.42, 4.93	
None of the time (5)	14	5.80	4.91, 6.70		5.06	3.94, 6.19	
SF-36v2® item 9g[c]				<0.001			<0.001
All of the time (1)	1	6.64	3.62, 9.66		5.74	2.00, 9.47	
Most of the time (2)	24	5.67	5.02, 6.32		4.79	3.99, 5.59	
Some of the time (3)	39	4.43	3.92, 4.93		3.97	3.35, 4.60	
A little of the time (4)	41	2.78	2.27, 3.29		2.24	1.60, 2.87	
None of the time (5)	36	2.07	1.54, 2.60		1.40	0.75, 2.06	
SF-36v2® item 9i[c]				<0.001			<0.001
All of the time (1)	5	8.00	6.69, 9.31		7.70	6.01, 9.38	
Most of the time (2)	36	5.26	4.73, 5.79		4.34	3.66, 5.03	

		NTDT-PRO T/W domain			NTDT-PRO SoB domain		
	n	LS mean	95% CI	*p* value[a]	LS mean	95% CI	*p* value[a]
Some of the time (3)	45	4.14	3.66, 4.61		3.58	2.97, 4.19	
A little of the time (4)	44	2.66	2.21, 3.11		2.08	1.50, 2.66	
None of the time (5)	11	1.21	0.35, 2.08		0.94	–0.18, 2.05	
PGI-S				<0.001			<0.001
0 to 2 (no symptoms)	40	1.33	0.95, 1.71		1.06	0.51, 1.60	
>2 to 4 (mild)	37	3.70	3.31, 4.10		2.83	2.27, 3.40	
>4 to 6 (moderate)	44	4.90	4.52, 5.29		4.08	3.53, 4.63	
>6 to 8 (severe)	19	5.75	5.21, 6.30		5.17	4.39, 5.96	
>8 (very severe)	5	7.70	6.67, 8.72		7.43	5.96, 8.91	

[a]*F*-test comparing T/W and SoB domain scores across subgroups (ANCOVA).
[b]"Please select one answer […] to indicate your response as it applies to the past 7 days": item HI12, "I feel weak all over"; item An2, "I feel tired".
[c]"How much of the time during the past week did you…": item 9e, "…have a lot of energy?"; item 9g, "…feel worn out?"; item 9i, "…feel tired?"
ANCOVA, analysis of covariance; CI, confidence interval; FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FS, Fatigue Subscale; LS, least-squares; PGI-S, Patient Global Impression of Severity; SF-36v2®, Short Form Health Survey version 2; SoB, Shortness of Breath; T/W, Tiredness/Weakness.

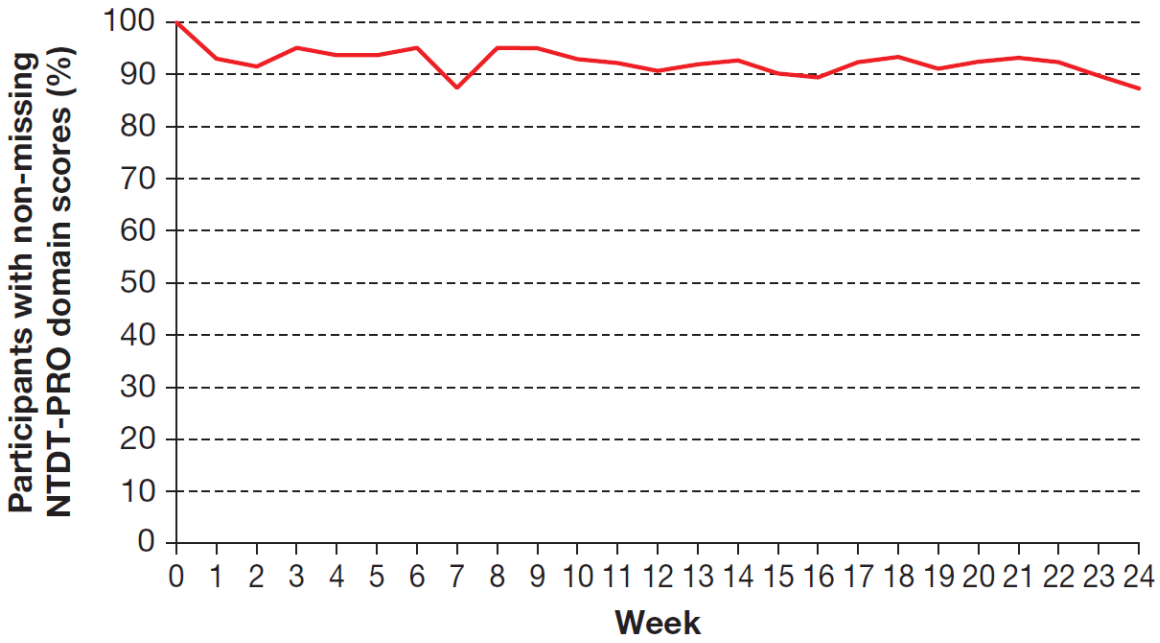

Figure S1 Percentage of participants with non-missing weekly NTDT-PRO domain scores**.** The percentage for a given week was calculated as the number of participants with non-missing weekly NTDT-PRO domain scores divided by the number of participants who remained on-study. For all weeks, percentages were the same for both the T/W and SoB domains.
SoB, Shortness of Breath; T/W, Tiredness/Weakness.

BMJ Open

Psychometric evaluation of the NTDT-PRO questionnaire for assessing symptoms in patients with non-transfusion-dependent beta-thalassaemia

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-066683.R1
Article Type:	Original research
Date Submitted by the Author:	25-Jan-2023
Complete List of Authors:	Taher, Ali T.; American University of Beirut Medical Center, Department of Internal Medicine Musallam, Khaled M.; Thalassemia Center; International Network of Hematology Viprakasit, Vip; Mahidol University, Division of Hematology & Oncology, Department of Pediatrics & Siriraj Thalassemia Center, Siriraj Research Hospital Kattamis, Antonis; National and Kapodistrian University of Athens, First Department of Pediatrics Lord-Bessen, Jennifer; Bristol Myers Squibb Co Yucel, Aylin; Bristol Myers Squibb Co Guo, Shien; Evidera Waltham Pelligra, Christopher; Evidera Shields, Alan L.; Adelphi Values Boston Shetty, Jeevan K.; Celgene International Sàrl Miteva, Dimana; Celgene International Sàrl Bueno, Luciana Moro; Celgene International Sàrl Cappellini , MD; University of Milan, Department of Internal Medicine, Fondazione IRCCS Ca' Granda Policlinico Hospital
Primary Subject Heading:	Haematology (incl blood transfusion)
Secondary Subject Heading:	Haematology (incl blood transfusion)
Keywords:	Anaemia < HAEMATOLOGY, Blood bank & transfusion medicine < HAEMATOLOGY, Clinical trials < THERAPEUTICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Psychometric evaluation of the NTDT-PRO questionnaire for assessing symptoms in patients with non-transfusion-dependent beta-thalassaemia

Ali T. Taher,[1] Khaled M. Musallam,[2,3] Vip Viprakasit,[4] Antonis Kattamis,[5] Jennifer Lord-Bessen,[6] Aylin Yucel,[6] Shien Guo,[7] Christopher Pelligra,[8] Alan L. Shields,[9] Jeevan K. Shetty,[10] Dimana Miteva,[10] Luciana Moro Bueno,[10] Maria Domenica Cappellini[11]

[1]Department of Internal Medicine, American University of Beirut Medical Center, Beirut, Lebanon

[2]Thalassemia Center, Burjeel Medical City, Abu Dhabi, UAE

[3]International Network of Hematology, London, UK

[4]Division of Hematology & Oncology, Department of Pediatrics & Siriraj Thalassemia Center, Siriraj Research Hospital, Mahidol University, Bangkok, Thailand

[5]First Department of Pediatrics, National and Kapodistrian University of Athens, Athens, Greece

[6]Bristol Myers Squibb, Princeton, NJ, USA

[7]Evidera, Waltham, MA, USA

[8]Evidera, Bogotá, Colombia

[9]Adelphi Values, Boston, MA, USA

[10]Celgene International Sàrl, a Bristol-Myers Squibb Company, Boudry, Switzerland

[11]Department of Internal Medicine, Fondazione IRCCS Ca' Granda Policlinico Hospital, University of Milan, Milan, Italy

ORCID

AT Taher: 0000-0001-8515-2238

KM Musallam: 0000-0003-3935-903X

V Viprakasit: 0000-0003-3162-1849

A Kattamis: 0000-0002-5178-0655

C Pelligra: 0000-0002-5255-2777

MD Cappellini: 0000-0001-8676-6864

Correspondence

Ali T. Taher, M.D., Ph.D., F.R.C.P.

American University of Beirut Medical Center, Halim and Aida Daniel Academic and Clinical Center, 4th floor, Hamra, Beirut, Lebanon

Telephone: +9611350000 Extension 5392

Email: ataher@aub.edu.lb

Keywords (3-6 words/phrases): psychometrics; non-transfusion-dependent beta-thalassaemia; patient-reported outcomes; symptom; anaemia

Running title: NTDT-PRO psychometric evaluation

Abstract

Objectives The NTDT-PRO questionnaire was developed for assessing anaemia-related Tiredness/Weakness (T/W) and Shortness of Breath (SoB) among patients with non-transfusion-dependent β-thalassaemia (NTDT). Psychometric properties were evaluated using blinded data from the BEYOND trial (NCT03342404).

Design Analysis of a phase 2, double-blind, randomised, placebo-controlled trial.

Setting USA, Greece, Italy, Lebanon, Thailand, and the UK.

Participants Adults (≥18 years) (N=145) with NTDT who had not received a red blood cell transfusion within 8 weeks prior to randomisation, with mean baseline haemoglobin level ≤10.0 g/dL.

Measures NTDT-PRO daily scores from baseline until week 24, and scores at select time points for the 36-Item Short Form Health Survey version 2 (SF-36v2®), Functional Assessment of Chronic Illness Therapy – Fatigue (FACIT-F), and Patient Global Impression of Severity (PGI-S).

Results Cronbach's alpha at weeks 13–24 was 0.95 and 0.84 for the T/W and SoB domains, respectively, indicating acceptable internal consistency reliability. Among participants self-reporting no change in thalassaemia symptoms via the PGI-S between baseline and week 1, intraclass correlation coefficients were 0.94 and 0.92 for the T/W and SoB domains, respectively, indicating excellent test–retest reliability. In a known-groups validity analysis, least-squares mean T/W and SoB scores at weeks 13–24 were worse in participants with worse scores for the FACIT-F Fatigue Subscale (FS), SF-36v2® vitality, or PGI-S. Indicating responsiveness, changes in T/W and SoB domain scores were moderately correlated with changes in haemoglobin levels, and strongly correlated with changes in SF-36v2® vitality, FACIT-F FS, select FACIT-F items, and the PGI-S. Improvements in least-squares mean T/W and SoB scores were higher in participants with greater improvements in scores on other patient-reported outcomes measuring similar constructs.

Conclusions The NTDT-PRO demonstrated adequate psychometric properties to assess anaemia-related symptoms in adults with NTDT and can be used to evaluate treatment efficacy in clinical trials.

Strengths and limitations of this study

- Strengths of this study include use of well-validated PRO instruments such as PGI-S, PGI-C, SF-36v2®, and FACIT-F.
- The data used in this analysis were from a phase 2 interventional study with participants from multiple geographic regions and spanning a range of NTDT symptom severities.
- The use of blinded data from an interventional study allowed for changes in symptom severity to be observed, validating the NTDT-PRO's sensitivity to identify longitudinal changes in symptoms.
- Given that NTDT is a rare disease, limitations of the present study include the reduced sample size for typical psychometric evaluations.
- Cut-off values used to define different levels of improvement in the responsiveness analysis are not well established and were based on certain assumptions.

INTRODUCTION

β-thalassaemias are a group of genetic blood disorders characterised by defective synthesis of the β-globin chains of haemoglobin and ineffective erythropoiesis. Phenotypes are highly variable: while some patients are borderline asymptomatic, others experience significant symptoms associated with severe chronic anaemia.[1]

From a clinical perspective, patients are often categorised as having transfusion-dependent β-thalassaemia (TDT) or non-transfusion-dependent β-thalassaemia (NTDT). While patients with TDT require lifelong blood transfusions, those with NTDT only require transfusions in certain circumstances, such as during infections, pregnancy, and surgery.[2,3] Due to anaemia or primary iron overload, which accumulate as patients get older, NTDT can result in various comorbidities (e.g., hepatic disease, endocrinopathy, thromboembolic events, pulmonary hypertension, leg ulcers, and extramedullary haematopoietic [EMH] masses), which not only have a negative impact on patients' daily activities and quality of life (QoL), but also reduce survival.[4-6]

Patient-reported outcome (PRO) questionnaires are used to assess how patients feel and function as well as their overall QoL. Reflecting the patient experience in these ways is important when evaluating treatments in clinical trials, and particularly in instances when patients experience symptoms from lifelong diseases.

Patient-centred research in NTDT is limited by a lack of rigorously developed PRO instruments for assessing symptoms important to patients in the target patient population. For example, health-related QoL (HRQoL) in patients with β-thalassaemias has typically been evaluated by generic questionnaires such as the Short Form Health Survey version 2 (SF-36v2®) and the World Health Organization 100-item Quality of Life Survey (WHOQOL-100),[7,8] which may fail to capture the unique experiences of patients with β-thalassaemia. Two β-thalassaemia-specific PRO instruments for assessing HRQoL are now available: the Specific Thalassaemia Quality of Life Instrument (STQOLI) and the Transfusion-dependent Quality of Life (TranQoL) questionnaire.[9,10] However, both tools were developed for patients with TDT and include questions on the impact of transfusions, which are often not relevant for patients with NTDT. Moreover, they focus more on general functioning and QoL and do not specifically capture anaemia-related symptoms of β-thalassaemia, which can be more

prominent in NTDT than in TDT because of the lack of transfusions.[11,12] In addition, neither instrument has been evaluated in patients with NTDT.

The NTDT-PRO was created to fill the gap in available, indication-specific PRO questionnaires defensible for use among patients with NTDT. Developed in the context of evaluating the treatment benefit of luspatercept (an approved treatment for anaemia in adults with TDT) among patients with NTDT, the NTDT-PRO is a 6-item questionnaire intended to measure the most relevant and important anaemia-related symptoms of NTDT.[13] In accordance with US Food and Drug Administration (FDA) guidance on the development of PRO tools,[14] evidence supporting the content validity of the NTDT-PRO was obtained from qualitative work, including concept elicitation and cognitive interviews with patients with NTDT,[13] and a preliminary psychometric evaluation using data from a 24-week observational study showed promising reliability and validity results.[15] However, the ability of the NTDT-PRO to capture longitudinal changes in symptoms could not be properly assessed due to the non-interventional study design. In the present study, a detailed evaluation of the reliability and validity of the NTDT-PRO was conducted, including its ability to reflect changes in symptom severity over time, using data from the BEYOND trial.[16]

METHODS

Study design

The analysis was based on blinded data generated from BEYOND, a phase 2, double-blind, randomised, placebo-controlled trial of luspatercept in adults with NTDT (NCT03342404), conducted in the USA, Greece, Italy, Lebanon, Thailand, and the UK.[16] Briefly, the trial included double-blind and open-label treatment phases and long-term follow-up. For double-blind treatment, participants were randomly assigned 2:1 to luspatercept or placebo. Luspatercept was administered as a subcutaneous injection every 3 weeks for 48 weeks. The assessment period for the primary and key secondary efficacy endpoints was weeks 13–24. The starting dose of luspatercept was 1 mg/kg and the maximum dose was 1.25 mg/kg or 120 mg. The trial was unblinded 48 weeks after the last participant had received their first dose of study drug. All participants were eligible to receive open-

label luspatercept for up to 15 months, and could then continue to receive luspatercept during the post-treatment follow-up period.

BEYOND received institutional review board/ethics committee approval and was conducted in accordance with International Council for Harmonisation Good Clinical Practice and the Declaration of Helsinki.

The psychometric analysis plan was finalized prior to the finalization of the core study statistical analysis plan and study unblinding. All analyses were carried out on an interim blinded datacut, and all analysts remained blinded until programming of all pre-specified analyses were complete.

Participants

Participants were adults (≥18 years of age) with β-thalassaemia or haemoglobin E/β-thalassaemia. They were non-transfusion-dependent, as defined by receipt of 0 to 5 units of red blood cells during the 24 weeks before randomisation, and had not received a red blood cell transfusion in the 8 weeks prior to randomisation. To be eligible for enrolment, they were additionally required to have a mean baseline haemoglobin level (based on at least 2 measurements taken ≥1 week apart) of ≤10.0 g/dL and an Eastern Cooperative Oncology Group (ECOG) performance status of 0 or 1. Patients with haemoglobin S/β-thalassaemia or α-thalassaemia alone were excluded, as were patients who had previously been exposed to luspatercept or sotatercept. All participants provided written informed consent.

Patient and public involvement

No patients involved.

PRO assessments

The NTDT-PRO and Patient Global Impression of Severity (PGI-S) were translated and linguistically validated into multiple languages based on the geographic regions of the study sites and were administered daily, in the preferred language of each participant, from the 7 days prior to randomisation until week 24, then daily for 7 days before dosing of every other dose of study drug.

The Patient Global Impression of Change (PGI-C), SF-36v2®, and Functional Assessment of Chronic Illness Therapy – Fatigue (FACIT-F) were administered at screening and on the day of dosing for every other dose of study drug, starting from the first dose. The SF-36v2®, FACIT-F, and PGI-C assessments were mapped to a nominal week using a mapping algorithm (see online supplementary table S1).

NTDT-PRO

The NTDT-PRO assesses the severity of symptoms associated with NTDT in the 24 hours prior to administration. The 6 items assess tiredness (lack of energy, 2 items), weakness (lack of strength, 2 items), and shortness of breath (2 items) when doing and when not doing physical activity. Each item uses an 11-point numeric rating scale (NRS) ranging from 0 (no symptoms) to 10 (extreme symptoms). Responses to the NTDT-PRO can be used to derive Tiredness/Weakness (T/W) and Shortness of Breath (SoB) domain scores. In the BEYOND trial, the NTDT-PRO was completed in the evening as a part of an electronic diary that also included the PGI-S. NTDT-PRO T/W and SoB scores were included as secondary endpoints in the trial.[16]

Weekly item and domain scores were calculated from baseline (week 0) to week 24. For a given week, the weekly score for each item was calculated as the average of the daily scores for that item if scores were available for at least 4 days (i.e., at least 50% of the week); otherwise, the score was set to "missing." Weekly T/W and SoB domain scores (range: 0 [no symptoms] to 10 [extreme symptoms]) were calculated as the average of non-missing weekly item scores for the T/W domain or SoB domain. Weekly domain scores were only calculated if weekly scores were non-missing for at least 2 of the 4 tiredness/weakness items (including ≥1 tiredness item and ≥1 weakness item) or at least 1 of the 2 shortness of breath items; otherwise, they were set to "missing." Average T/W and SoB scores over weeks 13–24 were calculated using data for all non-missing weeks during that time interval. If all weekly scores over weeks 13–24 were missing, the average score over weeks 13–24 was set to "missing".

PGI-S

PGI-S is a single-item questionnaire that assesses a patient's perception of their overall thalassaemia symptom severity in the previous 24 hours on an 11-point NRS ranging from 0 (no symptoms) to 10 (very severe symptoms). The weekly PGI-S score was calculated as the average of the daily scores if scores were available for at least 4 days; otherwise, it was set to "missing". Average PGI-S scores over weeks 13–24 were calculated using data for all non-missing weeks.

PGI-C

PGI-C is a single-item questionnaire that assesses a patient's perception of how their symptoms have changed over time. In BEYOND, participants responded to the question "How would you rate the overall change in your thalassaemia symptoms since the start of this study?" by selecting 1 of 7 response options ranging from "A great deal better" to "A great deal worse".

SF-36v2®

SF-36v2® consists of 8 multi-item scales assessing the following aspects of health over the previous 7 days: physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional, and mental health. SF-36v2® data were scored using Health Outcomes™ Scoring Software 5 (QualityMetric, Lincoln, RI, USA).[17] For each multi-item scale, the average of all items within the scale was calculated and the raw scores were converted to a 0 to 100 scale. They were then transformed to a US norm-based T-score (mean: 50, standard deviation [SD]: 10), with a higher T-score indicating better health. Finally, the Physical Component Summary (PCS) and Mental Component Summary (MCS) were derived as weighted averages of the T-scores for the 8 multi-item scales.

FACIT-F

FACIT-F is a 40-item questionnaire assessing fatigue and its effects on functioning and daily activities. It consists of the 27-item Functional Assessment of Cancer Therapy – General (FACT-G)

questionnaire and the 13-item Fatigue Subscale (FS). All items have a 7-day recall period and are rated on a 5-point scale ranging from "Not at all" to "Very much".

FACT-G comprises 4 domains: physical well-being (7 items, range: 0 to 28 points), social/family well-being (7 items, range: 0 to 28 points), emotional well-being (6 items, range: 0 to 24 points), and functional well-being (7 items, range: 0 to 28 points). Scores for each FACT-G domain and the FS (range: 0 to 52 points) were derived by summing the scores for the individual items (after reverse scoring, as applicable).[18]

Scores for 3 additional summary scales were also calculated: FACT-G total score=sum of scores for all FACT-G items (range: 0 to 108 points); FACIT-F trial outcome index (TOI)=sum of the scores for FACT-G physical well-being, FACT-G functional well-being, and the FS (range: 0 to 108 points); and FACIT-F total score=sum of scores for all FACT-G items and the FS (range: 0 to 160 points). For the FACT-G domains, the FS, and the additional summary scales, a higher score indicates less fatigue or better HRQoL.

Statistical analyses

All statistical analyses were conducted using SAS version 9.4 (SAS Institute Inc., Cary, NC, USA). Analyses were performed on blinded data collected up to week 24 during double-blind treatment (data cut-off: January 7, 2020) using the intent-to-treat (ITT) population, defined as all randomised participants. Summary statistics were calculated for demographics, baseline clinical characteristics, and PRO scores. For NTDT-PRO scores, floor and ceiling effects were also assessed.

Quality of completion of the NTDT-PRO was evaluated by calculating the percentages of participants with missing and non-missing weekly scores from among participants who were eligible for the assessment. Item–item and item–domain correlations for the NTDT-PRO were assessed by calculating Spearman's rank correlation coefficients, which were interpreted as <0.3=weak, ≥0.3 to <0.7=moderate, ≥0.7 to <0.9=strong, and ≥0.9=very strong.[19]

Confirmation of the weekly scoring rule

To evaluate whether modifying the weekly scoring rule for the NTDT-PRO would impact the variability of weekly item scores, an analysis was conducted at baseline, weeks 1, 2, 4, 8, 12, 16, 20, and 24, including data only from those participants with no missing daily item scores within each week. For each participant, a weekly score for each item was generated using a bootstrapping approach without replacement by randomly selecting a specific number of daily scores during the week according to the missing day scenario (scores missing for 1, 2, 3, 4, 5, or 6 days). For each missing-day scenario, each participant's simulated weekly item score was calculated as the mean of randomly selected daily scores. The average score across weeks was then calculated for each participant. Finally, the mean and SD were calculated across participants. To identify the point at which substantial changes in the variability of weekly item scores occurred, the SD for each missing-day scenario was compared with the SD when no days were missing using the Brown–Forsythe test.[20]

Reliability

Internal consistency reliability reflects the extent to which individual items from a scale consisting of multiple items are measuring the same general concept when measured at a single time point. In the present context, Cronbach's alpha[21] was calculated for weekly NTDT-PRO T/W and SoB domain scores with standardisation of variances before and after deletion of individual NTDT-PRO weekly items for the T/W domain score. Cronbach's alpha was deemed an appropriate measure of internal consistency for the NTDT-PRO T/W and SoB as previous exploratory factor analyses supported the grouping of the 4 tiredness and weakness items into 1 domain and the 2 shortness of breath items into another domain.[15] Values ≥0.70 indicated acceptable internal consistency.[22]

Test–retest reliability is a measure of how consistently an instrument measures a concept at different time points in "stable" participants, and was assessed, at the NTDT-PRO domain level, by calculating the intraclass correlation coefficient (ICC) for weekly domain scores using a 2-way mixed-effects analysis of variance (ANOVA) model with week as a fixed effect.[23] Stable

participants were those with PGI-S weekly scores at baseline and week 1 that differed by ≤0.5 points. An ICC of ≥0.70 indicated acceptable test–retest reliability.[24]

Validity

Convergent validity is demonstrated when different measures of the same concept are strongly correlated with each other, while discriminant validity can be inferred when unrelated concepts are weakly correlated. Convergent and discriminant validity was assessed via Spearman's rank correlation coefficients between NTDT-PRO domain scores and other scores (PGI-S score, and domain and summary scores for the SF-36v2® and FACIT-F) from assessments done at the same time point (baseline, week 24, or weeks 13–24). It was hypothesised that NTDT-PRO domain scores would be moderately to strongly related (Spearman's rank correlation coefficient: ≥0.3) to SF-36v2® physical functioning and vitality, FACIT-F physical well-being and FS, and the PGI-S scores, and less related (Spearman's rank correlation coefficient: <0.3) to SF-36v2® bodily pain, role-emotional, and MCS scores.

Known-groups validity of the NTDT-PRO domains—sensitivity to differentiate among groups of participants known to be clinically different—was assessed by comparing least-squares (LS) mean NTDT-PRO scores between different subgroups of participants, classified based on scores for the PGI-S, the FACIT-F FS, SF-36v2® vitality, and selected FACIT-F items and SF-36v2® items. The domains and items were selected for their theorised relationship to the concepts being measured by the NTDT-PRO T/W and SoB domains. Classifications used to define known groups are shown in online supplementary table S2. Classifications for the PGI-S were defined based on the assumption of a 2-point meaningful difference. For the FACIT-F FS, the cut-off used by the instrument developer to differentiate patients with cancer from the general population was used to classify participants as moderate or mild.[25] A clinically important difference of 3 points, as suggested by instrument developer, was used to define the other categories.[26] The SF-36v2® vitality "normal" category was defined based on a meaningful difference of ±6.7 points from the norm-based mean score of 50, with other categories defined by subsequently adding or subtracting 6.7 from the upper or lower bounds, respectively.[17] For item-based known groups, each verbal response level was taken as a known

group. Analysis of covariance (ANCOVA) models were used that included NTDT-PRO domain scores at baseline, week 24, and weeks 13–24 as the dependent variable, and the known-groups measure at the corresponding time point as the independent variable, and that were adjusted for age and geographic region.

Responsiveness

Responsiveness was defined as the sensitivity of the NTDT-PRO to changes in a patient's symptom severity over time. Responsiveness was evaluated by first calculating Spearman's rank correlation coefficients for changes from baseline in NTDT-PRO domain scores at week 24 and weeks 13–24 and the changes in haemoglobin level (generally considered as a measure of response) and scores for FACIT-F FS, SF-36v2® vitality, the PGI-S, the PGI-C, and selected FACIT-F and SF-36v2® items. The 5 measures with the strongest correlations at weeks 13–24 with NTDT-PRO domain score changes were included in a subsequent analysis where ANCOVA models were used to compare LS mean changes in NTDT-PRO domain scores among different response categories. Response categories (table 1) were defined based on reported estimates of clinically meaningful within-patient changes for FACIT-F FS and SF-36v2® vitality domain scores or 1-point differences for individual items. A 1-point difference was also used to define the response categories of the PGI-S. The models included NTDT-PRO domain scores change as the dependent variable and response categories for the given anchor measure as the independent variable, and were adjusted for age and geographic region.

Table 1 Responsiveness at weeks 13–24

	Spearman's rank correlation coefficient (r)[a]		LS mean change (95% CI) at weeks 13–24[b]				
	Week 24	Weeks 13–24	Improvement level 2	Improvement level 1	No change	Worsening	*p* value[c]
NTDT-PRO T/W domain							
Haemoglobin level	–0.38	–0.30	–	–	–	–	–
SF-36v2® vitality	–0.49	–0.46	–	–1.77 (–2.42, –1.12)	–0.40 (–0.80, 0.00)	0.60 (–0.20, 1.39)	<0.001
SF-36v2® item 9e	0.28	0.41	–	–	–	–	–
SF-36v2® item 9g	–0.41	–0.40	–	–	–	–	–
SF-36v2® item 9i	–0.42	–0.43	–	–	–	–	–
FACIT-F FS	–0.52	–0.56	–2.74 (–3./42, –2.06)	–1.68 (–2.44, –0.93)	–0.22 (–0.57, 0.13)	0.42 (–0.16, 1.01)	<0.001
FACIT-F item HI7	–0.41	–0.40	–	–	–	–	–
FACIT-F item HI12	–0.58	–0.60	–3.28 (–4.24, –2.32)	–1.69 (–2.44, –0.95)	–0.51 (–0.88, –0.13)	0.48 (–0.08, 1.03)	<0.001
FACIT-F item An2	–0.43	–0.45	–	–1.84 (–2.46, –1.22)	–0.21 (–0.61, 0.20)	0.00 (–0.68, 0.68)	<0.001
FACIT-F item An5	–0.33	–0.31	–	–	–	–	–
PGI-S	0.83	0.79	–3.26 (–3.75, –2.77)	–1.80 (–2.35, –1.25)	–0.09 (–0.35, 0.18)	0.99 (0.56, 1.42)	<0.001
PGI-C	0.39	0.28	–	–	–	–	–
NTDT-PRO SoB domain							
Haemoglobin level	–0.36	–0.32	–	–	–	–	–
SF-36v2® vitality	–0.40	–0.41	–	–1.28 (–1.91, –0.66)	–0.22 (–0.60, 0.16)	0.52 (–0.24, 1.28)	<0.001
SF-36v2® item 9e	0.30	0.41	–	–	–	–	–
SF-36v2® item 9g	–0.38	–0.36	–	–	–	–	–
SF-36v2® item 9i	–0.30	–0.34	–	–	–	–	–
FACIT-F FS	–0.49	–0.51	–2.21 (–2.88, –1.53)	–1.18 (–1.92, –0.43)	–0.01 (–0.36, 0.33)	0.25 (–0.32, 0.83)	<0.001
FACIT-F item HI7	–0.32	–0.29	–	–	–	–	–
FACIT-F item HI12	–0.45	–0.48	–2.70 (–3.64, –1.76)	–1.08 (–1.81, –0.35)	–0.25 (–0.62, 0.12)	0.33 (–0.22, 0.87)	<0.001

FACIT-F item An2	–0.39	–0.43	–	–1.38 (–1.97, –0.78)	–0.07 (–0.45, 0.32)	0.09 (–0.56, 0.74)	<0.001
FACIT-F item An5	–0.36	–0.31	–	–	–	–	–
PGI-S	0.68	0.69	–2.62 (–3.14, –2.09)	–1.17 (–1.77, –0.58)	0.00 (–0.28, 0.28)	1.01 (0.55, 1.47)	<0.001
PGI-C	0.30	0.28	–	–	–	–	–

[a]Changes from baseline.
[b]Score changes defining response categories (improvement level 2, improvement level 1, no change, worsening): SF-36v2® vitality: N/A, ≥6.7, >–6.7 to <6.7, ≤–6.7; FACIT-F FS: ≥8, 4 to <8, >–4 to <4, ≤–4; FACIT-F item HI12: ≥2, 1 to <2, >–1 to <1, ≤–1; FACIT-F item An2: N/A, ≥1, >–1 to <1, ≤–1; PGI-S: ≤–2, >–2 to –1, >–1 to <1, ≥1. For SF-36v2® vitality and FACIT-F item An2, no improvement level 2 category was used.
[c]*F*-test comparing T/W and SoB domain scores across response categories (ANCOVA).
ANCOVA, analysis of covariance; CI, confidence interval; FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FS, Fatigue Subscale; LS, least squares; N/A, not applicable; PGI-C, Patient Global Impression of Change; PGI-S, Patient Global Impression of Severity; SF-36v2 ®, Short Form Health Survey version 2; SoB, Shortness of Breath; T/W, Tiredness/Weakness.

RESULTS

Participants

The ITT population comprised 145 participants with a mean (SD) age of 39.9 (12.8) years (range: 18 to 71 years) (see online supplementary table S3). Most participants were female (56.6%), White (60.0%), and from North America or Europe (62.1%). A total of 26.9% of participants had a diagnosis of haemoglobin E/β-thalassaemia, and 6.2% had a diagnosis of β-thalassaemia combined with α-thalassaemia. The mean (SD) haemoglobin level at baseline was 8.2 (1.2) g/dL, and most participants had no or only a slight transfusion burden (mean: 0.3 units of red blood cells in the 24 weeks before the first dose of study drug). Most participants (69.0%) had an ECOG performance status of 0, indicating normal functioning.

Quality of completion of the NTDT-PRO

Across all NTDT-PRO items, the percentage of participants with <4 days of missing NTDT-PRO data (i.e., with sufficient data to calculate average weekly item scores) was 98.6% at baseline and 84.4% at week 24 (see online supplementary table S4). Across the first 24 weeks of treatment, at least 87.3% of participants per week had non-missing NTDT-PRO T/W and SoB scores (see online supplementary figure S1).

PRO score distributions at baseline

Average weekly NTDT-PRO item scores at baseline ranged from 2.4 for item 5-SobNA (shortness of breath not doing physical activity) to 5.0 for item 2-TiredPA (tiredness doing physical activity) (see online supplementary table S5). Baseline average weekly domain scores were 4.1 for T/W and 3.3 for SoB. The weekly average PGI-S score at baseline was 3.7, and average scores for the SF-36v2® scales and component summaries ranged from 42.2 for general health to 51.5 for bodily pain. The average baseline FACIT-F FS score of 36.4 was worse than that in the US general population (43.6).[24] Nonetheless, these data collectively suggested that participants generally had mild to moderate symptoms at study baseline.

Based on skewness and kurtosis values, the distributions of weekly T/W and SoB scores at baseline were generally symmetric but slightly platykurtic, indicating that few participants had extreme values. For T/W, 1.4% of participants had a score of 0 and 1.4% had a score >9; 7.6% of participants had an SoB score of 0 and 0.7% had an SoB score >9 (see online supplementary table S5). For each week up to week 24, <6% of participants had a T/W score of 0, <2% had a T/W score >9, <15% had an SoB score of 0, and <1% had an SoB score >9. This indicates that there were no problematic floor or ceiling effects.

NTDT-PRO item–item and item–domain correlations

Across the 3 assessment time points/time intervals, item 1-TiredNA (tiredness not doing physical activity) was very strongly correlated with item 3-WeakNA (weakness not doing physical activity) (r=0.97 to 0.98), and item 2-TiredPA was very strongly correlated with item 4-WeakPA (weakness doing physical activity) (r=0.98 to 0.99). Item 5-SobNA and item 6-SobPA (shortness of breath doing physical activity) were strongly correlated with each other (r=0.74 to 0.81) and moderately to strongly correlated with item 1-TiredNA, item 2-TiredPA, item 3-WeakNA, and item 4-WeakPA (r=0.50 to 0.81) (table 2).

At the domain level, T/W and SoB scores were strongly correlated with each other (r=0.77 to 0.79). As anticipated, item 1-TiredNA, item 2-TiredPA, item 3-WeakNA, and item 4-WeakPA correlated more strongly with T/W (r=0.88 to 0.95) than with SoB (r=0.67 to 0.77), and item 5-SobNA and item 6-SobPA correlated more strongly with SoB (r=0.89 to 0.97) than with T/W (r=0.64 to 0.78).

Table 2 NTDT-PRO item–item and item–domain correlations

	Spearman's rank correlation coefficient (r)							
	Item 1-TiredNA	**Item 2-TiredPA**	**Item 3-WeakNA**	**Item 4-WeakPA**	**Item 5-SobNA**	**Item-6 SobPA**	**T/W domain**	**SoB domain**
Baseline (N=145)								
Item 1-TiredNA	–	0.77	0.97	0.75	0.75	0.67	0.93	0.75
Item 2-TiredPA	0.77	–	0.73	0.98	0.57	0.77	0.94	0.72

	Spearman's rank correlation coefficient (r)							
	Item 1-TiredNA	**Item 2-TiredPA**	**Item 3-WeakNA**	**Item 4-WeakPA**	**Item 5-SobNA**	**Item-6 SobPA**	**T/W domain**	**SoB domain**
Item 3-WeakNA	0.97	0.73	–	0.74	0.77	0.65	0.91	0.74
Item 4-WeakPA	0.75	0.98	0.74	–	0.58	0.78	0.94	0.73
Item 5-SobNA	0.75	0.57	0.77	0.58	–	0.81	0.70	0.93
Item 6-SobPA	0.67	0.77	0.65	0.78	0.81	–	0.77	0.96
T/W domain	0.93	0.94	0.91	0.94	0.70	0.77	–	0.78
SoB domain	0.75	0.72	0.74	0.73	0.93	0.96	0.78	–
Week 24 (N=110)								
Item 1-TiredNA	–	0.73	0.97	0.71	0.76	0.59	0.89	0.69
Item 2-TiredPA	0.73	–	0.72	0.99	0.54	0.80	0.95	0.75
Item 3-WeakNA	0.97	0.72	–	0.72	0.80	0.62	0.89	0.73
Item 4-WeakPA	0.71	0.99	0.72	–	0.56	0.81	0.95	0.77
Item 5-SobNA	0.76	0.54	0.80	0.56	–	0.75	0.68	0.89
Item 6-SobPA	0.59	0.80	0.62	0.81	0.75	–	0.78	0.97
T/W domain	0.89	0.95	0.89	0.95	0.68	0.78	–	0.79
SoB domain	0.69	0.75	0.73	0.77	0.89	0.97	0.79	–
Weeks 13–24 (N=131)								
Item 1-TiredNA	–	0.71	0.98	0.70	0.73	0.57	0.88	0.67
Item 2-TiredPA	0.71	–	0.71	0.99	0.50	0.79	0.95	0.74
Item 3-WeakNA	0.98	0.71	–	0.72	0.77	0.61	0.89	0.72
Item 4-WeakPA	0.70	0.99	0.72	–	0.52	0.81	0.95	0.76
Item 5-SobNA	0.73	0.50	0.77	0.52	–	0.74	0.64	0.89
Item 6-SobPA	0.57	0.79	0.61	0.81	0.74	–	0.76	0.96
T/W domain	0.88	0.95	0.89	0.95	0.64	0.76	–	0.77
SoB domain	0.67	0.74	0.72	0.76	0.89	0.96	0.77	–

NTDT-PRO, non-transfusion-dependent β-thalassaemia-patient-reported outcomes; SoB, Shortness of Breath; SobNA, shortness of breath not doing physical activity; SobPA, shortness of breath doing physical activity; TiredNA, tiredness not doing physical activity; TiredPA, tiredness doing physical activity; WeakNA, weakness not doing physical activity; WeakPA, weakness doing physical activity; T/W, Tiredness/Weakness.

Weekly scoring rule

For all NTDT-PRO items, mean scores varied very little between different scenarios where the number of missing days ranged from 0 to 6 (see online supplementary table S6). Moreover, when comparing SD values for the different missing day scenarios using the Browne–Forsythe test, none of the SDs from the missing days were statistically significantly different from the SD when no days were missing. The requirement that scores be available for at least 4 days for a weekly score to be calculated was therefore shown to be reasonable.

Reliability

Internal consistency reliability

Cronbach's alpha for the NTDT-PRO T/W domain was 0.94 to 0.95 across the 3 assessment time points/time intervals (baseline, week 24, weeks 13–24) (see online supplementary table S7), indicating acceptable internal consistency reliability but suggesting possible item redundancy. However, removing individual items from the T/W domain did not increase Cronbach's alpha, indicating that there was no item redundancy. Cronbach's alpha for the NTDT-PRO SoB domain was 0.84 to 0.89, also indicating acceptable internal consistency reliability.

Test–retest reliability

In stable participants (those with a difference in PGI-S weekly scores of ≤0.5 points between baseline and week 1: N=73), ICC was 0.94 for the T/W domain and 0.92 for the SoB domain. These values were comfortably above the prespecified acceptability threshold of 0.70, indicating very good test–retest reliability.

Validity

Convergent and discriminant validity

Hypothesised convergent validity of NTDT-PRO with SF-36v2® physical functioning and vitality, FACIT-F physical well-being, FACIT-F FS, and PGI-S was demonstrated, with all correlation

coefficients exceeding the prespecified threshold of 0.3 in the expected direction (negative for the SF-36v2® and FACIT-F domains and positive for the PGI-S) (table 3). By contrast, with the exception of the weak correlation between SoB and SF-36v2® bodily pain at week 24 (r=–0.29), the hypothesised discriminant validity with SF-36v2® bodily pain, role-emotional, and MCS was not demonstrated.

Table 3 Convergent and discriminant validity

	Spearman's rank correlation coefficient (r)					
	NTDT-PRO T/W domain			NTDT-PRO SoB domain		
	Baseline	Week 24	Weeks 13–24	Baseline	Week 24	Weeks 13–24
SF-36v2®[a]						
Physical functioning	–0.50	–0.35	–0.43	–0.50	–0.35	–0.40
Role-physical	–0.65	–0.44	–0.50	–0.60	–0.40	–0.52
Bodily pain	–0.43	–0.34	–0.41	–0.38	–0.29	–0.37
General health	–0.53	–0.29	–0.34	–0.45	–0.37	–0.36
Vitality	–0.73	–0.61	–0.60	–0.61	–0.56	–0.52
Social functioning	–0.56	–0.34	–0.37	–0.55	–0.32	–0.44
Role-emotional	–0.55	–0.36	–0.43	–0.54	–0.31	–0.47
Mental health	–0.53	–0.38	–0.44	–0.50	–0.37	–0.43
PCS	–0.60	–0.35	–0.44	–0.54	–0.36	–0.43
MCS	–0.62	–0.46	–0.48	–0.58	–0.41	–0.47
FACIT-F[b]						
Physical well-being	–0.69	–0.55	–0.60	–0.60	–0.47	–0.51
Social/family well-being	–0.33	–0.27	–0.23	–0.30	–0.28	–0.22
Emotional well-being	–0.54	–0.35	–0.39	–0.50	–0.40	–0.41
Functional well-being	–0.62	–0.38	–0.42	–0.60	–0.44	–0.39
FACT-G total score	–0.66	–0.46	–0.49	–0.61	–0.47	–0.46
FACIT-F FS	–0.76	–0.58	–0.65	–0.66	–0.55	–0.52
FACIT-F TOI	–0.78	–0.55	–0.64	–0.69	–0.54	–0.54
FACIT-F total score	–0.74	–0.53	–0.58	–0.67	–0.52	–0.51
PGI-S[c]	0.86	0.83	0.80	0.72	0.67	0.65

[a]n=141 at baseline, n=96 at week 24, n=125 at weeks 13–24.
[b]n=144 at baseline, n=96 at week 24, n=126 at weeks 13–24.
[c]n=145 at baseline, n=110 at week 24, n=131 at weeks 13–24.
FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FACT-G, Functional Assessment of Cancer Therapy – General; FS, Fatigue Subscale; MCS, Mental Component Summary; NTDT-PRO, non-transfusion-dependent β-thalassaemia-patient-reported outcomes; PCS, Physical Component Summary; PGI-S, Patient Global Impression of Severity; SF-36v2®, Short Form Health Survey version 2; SoB, Shortness of Breath; TOI, trial outcome index; T/W, Tiredness/Weakness.

Known-groups validity

Known-groups validity was assessed using FACIT-F FS, SF-36v2® vitality, selected FACIT-F and SF-36v2® items, and the PGI-S. The FACIT-F and SF-36v2® items respectively measure similar concepts as the FACIT-F FS and SF-36v2® vitality but had the advantage of clearly defined rating scales that provided clear cut-off values to differentiate levels of severity. At weeks 13–24 (table 4), as well as at baseline (see online supplementary table S8) and week 24 (see online supplementary table S2), LS mean T/W and SoB scores on the NTDT-PRO were significantly higher (worse) in participants with lower (worse) scores for the FACIT-F FS, FACIT-F items HI12 (feeling weak all over) and An2 (feeling tired), SF-36v2® vitality, and SF-36v2® items 9g (feeling worn out) and 9i (feeling tired), and in participants with higher (worse) scores for SF-36v2® item 9e (having a lot of energy) and the PGI-S. Known-groups validity of the T/W and SoB domains was therefore demonstrated.

Table 4 Known-groups validity at weeks 13–24

		NTDT-PRO T/W domain			NTDT-PRO SoB domain		
	n	**LS mean**	**95% CI**	***p* value[a]**	**LS mean**	**95% CI**	***p* value[a]**
FACIT-F FS				<0.001			<0.001
Very severe (≤37)	43	4.39	3.90, 4.88		3.90	3.35, 4.45	
Severe (>37 to 40)	16	2.91	2.10, 3.73		1.77	0.86, 2.68	
Moderate (>40 to 43)	19	2.81	2.06, 3.55		2.61	1.77, 3.45	
Mild (>43 to 46)	17	1.86	1.05, 2.67		1.92	1.01, 2.83	
Very mild/no symptoms (>46)	31	1.17	0.57, 1.78		0.87	0.19, 1.55	
FACIT-F item HI12[b]				<0.001			<0.001
Very much (0)	5	5.50	4.08, 6.92		3.23	1.60, 4.87	
Quite a bit (1)	16	4.81	4.01, 5.60		4.26	3.34, 5.17	
Somewhat (2)	25	3.70	3.08, 4.33		3.51	2.79, 4.23	
A little bit (3)	53	2.57	2.08, 3.07		2.12	1.55, 2.68	
Not at all (4)	27	1.13	0.48, 1.79		0.84	0.09, 1.59	
FACIT-F item An2[b]				<0.001			<0.001
Very much (0)	8	5.33	4.10, 6.56		3.44	2.07, 4.81	
Quite a bit (1)	12	4.80	3.81, 5.80		4.18	3.08, 5.29	
Somewhat (2)	25	3.38	2.70, 4.07		3.55	2.78, 4.31	
A little bit (3)	64	2.44	1.94, 2.94		1.93	1.37, 2.48	
Not at all (4)	17	1.52	0.66, 2.38		1.20	0.25, 2.16	
SF-36v2® vitality				<0.001			<0.001
Very poor (≤36.6)	20	5.35	4.45, 6.26		4.54	3.54, 5.55	
Poor (>36.6 to 43.3)	19	4.51	3.54, 5.48		3.83	2.76, 4.89	
Normal (>43.3 to 56.7)	64	3.05	2.55, 3.55		2.82	2.27, 3.37	
Better (>56.7 to 63.4)	25	1.86	1.29, 2.44		1.34	0.70, 1.98	

		NTDT-PRO T/W domain			NTDT-PRO SoB domain		
	n	LS mean	95% CI	*p* value[a]	LS mean	95% CI	*p* value[a]
Much better (>63.4)	13	2.45	1.17, 3.73		2.14	0.72, 3.55	
SF-36v2® item 9e[c]				<0.001			<0.001
All of the time (1)	8	2.50	1.29, 3.71		1.69	0.32, 3.06	
Most of the time (2)	44	1.82	1.27, 2.36		1.69	1.07, 2.31	
Some of the time (3)	45	3.18	2.66, 3.70		2.65	2.06, 3.24	
A little of the time (4)	22	4.62	3.87, 5.37		4.43	3.58, 5.28	
None of the time (5)	6	5.64	4.28, 7.01		3.69	2.13, 5.24	
SF-36v2® item 9g[c]				<0.001			<0.001
All of the time (1)	4	5.92	4.30, 7.54		4.37	2.56, 6.19	
Most of the time (2)	11	5.30	4.31, 6.29		4.43	3.32, 5.53	
Some of the time (3)	34	3.49	2.93, 4.06		3.17	2.54, 3.80	
A little of the time (4)	49	2.67	2.16, 3.19		2.45	1.87, 3.03	
None of the time (5)	27	1.43	0.77, 2.09		0.83	0.09, 1.56	
SF-36v2® item 9i[c]				<0.001			<0.001
All of the time (1)	7	5.37	4.01, 6.73		4.01	2.51, 5.51	
Most of the time (2)	25	4.32	3.60, 5.05		3.88	3.08, 4.68	
Some of the time (3)	38	2.88	2.29, 3.47		2.55	1.90, 3.20	
A little of the time (4)	49	2.17	1.62, 2.73		1.72	1.11, 2.34	
None of the time (5)	6	2.21	0.76, 3.67		2.14	0.53, 3.74	
PGI-S				<0.001			<0.001
0 to 2 (no symptoms)	45	1.37	0.94, 1.79		1.10	0.57, 1.62	
>2 to 4 (mild)	36	2.93	2.47, 3.40		2.68	2.10, 3.26	
>4 to 6 (moderate)	34	4.48	3.99, 4.98		3.95	3.32, 4.57	
>6 to 8 (severe)	11	4.94	4.16, 5.73		4.18	3.20, 5.17	
>8 (very severe)	5	6.82	5.65, 7.98		5.91	4.45, 7.38	

[a]*F*-test comparing T/W and SoB domain scores across subgroups (ANCOVA).
[b]"Please select 1 answer […] to indicate your response as it applies to the past 7 days": item HI12, "I feel weak all over"; item An2, "I feel tired".
[c]"How much of the time during the past week did you…": item 9e, "…have a lot of energy?"; item 9g, "…feel worn out?"; item 9i, "…feel tired?"
ANCOVA, analysis of covariance; CI, confidence interval; FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FS, Fatigue Subscale; LS, least squares; NTDT-PRO, non-transfusion-dependent β-thalassaemia-patient-reported outcomes; PGI-S, Patient Global Impression of Severity; SF-36v2®, Short Form Health Survey version 2; SoB, Shortness of Breath; T/W, Tiredness/Weakness.

Responsiveness

Considering changes from baseline to week 24 and weeks 13–24, NTDT-PRO T/W and SoB domain scores were moderately correlated with changes in haemoglobin level (–0.30 to –0.38) and weakly to moderately correlated with the PGI-C (0.28 to 0.39) (table 1). The strongest correlations for the T/W and SoB domain score changes were with changes on SF-36v2® vitality (–0.40 to –0.49), the FACIT-F FS (–0.49 to –0.56), FACIT-F items HI12 (feeling weak all over, –0.45 to –0.60) and An2 (feeling tired, –0.39 to –0.45), and the PGI-S (0. 68 to 0.83). In a responsiveness analysis using these 5 measures as anchors, decreases (improvements) in LS mean T/W and SoB scores were significantly

higher in participants with greater improvements in scores on the anchors. The T/W and SoB domains were therefore shown to be responsive to changes in symptom severity (table 1).

DISCUSSION

Broadly, the NTDT-PRO demonstrated sufficient psychometric performance to defend its use as a measure of treatment outcome in clinical research among patients with NTDT. Distributional properties were good, as illustrated by the lack of floor and ceiling effects. High ICC values in patients assessed as stable based on PGI-S scores at baseline and week 1 indicated good test–retest reliability, while similarly high Cronbach's alpha coefficients at baseline, week 24, and weeks 13–24 indicated good internal consistency reliability. Correlation analyses confirmed the hypothesised direction and strength of relationship of both NTDT-PRO domains with other PRO measures, although the hypothesised discriminant validity with SF-36v2® bodily pain, role-emotional, and MCS was not demonstrated. However, as weakness, tiredness, and shortness of breath are broad concepts, it was not wholly surprising that NTDT-PRO T/W and SoB domain scores were correlated with these SF-36v2® scores. Finally, known-groups validity and responsiveness were demonstrated based on the PGI-S and selected FACIT-F and SF-36v2® items.

These findings build on an earlier preliminary psychometric analysis using data from 48 adults with NTDT who participated in a multicentre observational study, which demonstrated that the NTDT-PRO had high internal consistency reliability and test–retest reliability.[15] That earlier study was unable to adequately evaluate sensitivity to change, however, due to its non-interventional study design. This resulted in very few participants experiencing improvement in symptoms, as assessed by the PGI-C. In the present analysis, using data from the first 24 weeks of treatment in the BEYOND trial, the relationship among changes in NTDT-PRO scores relative to changes observed in multiple other measures of similar and distinct concepts at week 24 and weeks 13–24 were as we hypothesised, and are supportive of the tool's ability to detect change.

Although the NTDT-PRO T/W and SoB domains were shown to be responsive to changes over time on all the anchors examined in the responsiveness analysis, PGI-C scores had the weakest correlation (0.28) with change in T/W domain score at weeks 13–24 among the included anchors. The

weaker correlation between the NTDT-PRO domain score changes and the PGI-C as compared to other potential anchors may be due to an issue with recall: it may have been difficult for patients to rate how much their overall thalassaemia symptoms—which can be many—had changed in the 24 weeks since the beginning of the study.[27,28]

Limitations of the present study include the modest sample size for typical psychometric evaluations, although it was adequate for assessment of the trial endpoints. NTDT is a rare disease, which makes recruitment challenging. Moreover, cut-off values defining different levels of improvement are not yet well established for some of the anchors included in the responsiveness analysis (PGI-S, FACIT-F FS, and SF-36v2® vitality), so the cut-off values used in the responsiveness analysis were necessarily based on certain assumptions. However, given that score changes for these PRO measures were moderately to strongly correlated with score changes for the NTDT-PRO domains, modifying the cut-off values used to define different levels of improvement would likely yield very similar findings. Strengths of this study include use of well-validated PRO instruments, including the SF-36v2® and FACIT-F. Additionally, data for this analysis were from a phase 2 interventional study with participants from multiple geographic regions and spanning a range of NTDT symptom severities based on baseline T/W and SoB domain scores. This confirms the validity of the NTDT-PRO over a broad population. The use of data from an interventional study also allowed for changes in symptom severity to be observed, validating the sensitivity of the NTDT-PRO to changes in symptoms.

In conclusion, the NTDT-PRO demonstrated adequate reliability, validity, and responsiveness when used to assess tiredness/weakness and shortness of breath in patients with NTDT. As a fully validated PRO instrument, it can be used to confidently assess the efficacy of treatments targeting anaemia in clinical studies for NTDT. The instrument was developed for research purposes and to inform trial endpoints, but its practical use in the clinical setting warrants further evaluation. Future analyses will focus on the NTDT-PRO score interpretability by identifying meaningful change thresholds and symptomatic thresholds for the T/W and SoB domains.

Acknowledgments

The authors received medical writing support in the preparation of this manuscript from Stephen Gilliver, PhD, of Evidera, and editorial support from Patricia Fonseca, PhD, of Excerpta Medica, funded by Bristol Myers Squibb.

Data availability statement

The data that support the findings of this study are available in the supplementary material of this article and from the corresponding author.

Contributors

JL-B, SG, AY, CP, and ALS contributed to protocol development. SG, CP, and ALS made substantial contributions to the design and concept of the study. ATT, VV, AK, and MDC contributed to data acquisition. SG and CP conducted the data and statistical analysis. ATT, KMM, VV, AK, JL-B, AY, SG, CP, ALS, JKS, DM, LMB, and MDC interpreted the data, revised the work for intellectual content, provided final approval of the version to be published, and agree to be accountable for all aspects of the work related to accuracy and integrity. ATT accepts responsibility for the overall content as the guarantor. The guarantor accepts full responsibility for the finished work and/or conduct of the study, had access to the data, and controlled the decision to publish.

Funding

This study was funded by Bristol Myers Squibb. Award/grant number: not applicable.

Competing interests

ATT: consulting fees from Agios Pharmaceuticals; research funding and consulting fees from Celgene/Bristol Myers Squibb, Ionis Pharmaceuticals, Novartis Pharmaceuticals, and Vifor Pharma. KMM: consulting fees from Agios Pharmaceuticals, Celgene/Bristol Myers Squibb, CRISPR Therapeutics, Novartis, Pharmacosmos, and Vifor Pharma. VV: research funding from Bristol Myers Squibb. AK: advisory board fees and consulting fees from Agios Pharmaceuticals, Celgene/Bristol

Myers Squibb, Chiesi Farmaceutici, CRISPR Therapeutics/Vertex Pharmaceuticals, Ionis Pharmaceuticals, Novartis, and Vifor Pharma; research support from Celgene/Bristol Myers Squibb and Novartis. JL-B, AY, JKS, and LMB: employment by and stock/equity holder of Bristol Myers Squibb. SG: employment by Evidera; consultancy fees from Bristol Myers Squibb, Gilead, and Janssen. CP: employment by Evidera. ALS: employment by Adelphi Values. DM: employment by Bristol Myers Squibb. MDC: advisory board fees from Celgene/Bristol Myers Squibb, CRISPR Therapeutics, Ionis Pharmaceuticals, Novartis, Novo Nordisk, Sanofi Genzyme, and Vifor Pharma.

Ethics approval

The BEYOND trial received institutional review board/ethics committee approval (sites 101 and 102, A Kattamis and E Voskaridou: 112/17; site 201, MD Cappellini: CE150176; site 202: GL Forni: CE150176 and CE150124; site 203, S Perrotta: CE150176 and CE150110; site 204, AG Piga: CE150176 and CE150089; site 206, A Filosa: CE150176 and CE150040; site 301, AT Taher: NA and BIO-2017-0338; site 401: V Viprakasit: 689/2560(EC4); site 501, TD Coates: CHLA-17-00444; site 503, AA Thompson: IRB 2018-1580; and site 601, JB Porter: 17/EM/0438) and was conducted in accordance with International Council for Harmonisation Good Clinical Practice and the Declaration of Helsinki.

Patient consent for publication

Not required for this analysis.

Clinical trial registration

ClinicalTrials.gov Identifier: NCT03342404 (BEYOND)

REFERENCES

1 Taher AT, Musallam KM, Cappellini MD. Beta-thalassemias. *N Engl J Med* 2021;384:727–43.
2 Musallam KM, Rivella S, Vichinsky E, *et al*. Non-transfusion-dependent thalassemias. *Haematologica.* 2013;98(6):833-844.
3 Musallam KM, Cappellini MD, Viprakasit V, *et al*. Revisiting the non-transfusion-dependent (NTDT) vs. transfusion-dependent (TDT) thalassemia classification 10 years later. *Am J Hematol* 2021;96:E54–6.
4 Taher AT, Musallam KM, El-Beshlawy A, *et al*. Age-related complications in treatment-naive patients with thalassaemia intermedia. *Br J Haematol* 2010;150:486–9.
5 Musallam KM, Vitrano A, Meloni A, *et al*. Survival and causes of death in 2,033 patients with non-transfusion-dependent β-thalassemia. *Haematologica.* 2021;106:2489–92.
6 Musallam KM, Vitrano A, Meloni A, *et al*. Risk of mortality from anemia and iron overload in nontransfusion-dependent β-thalassemia. *Am J Hematol* 2022;97:E78–80.
7 Arian M, Mirmohammadkhani M, Ghorbani R, *et al*. Health-related quality of life (HRQoL) in beta-thalassemia major (β-TM) patients assessed by 36-item short form health survey (SF-36): a meta-analysis. *Qual Life Res* 2019;28:321–34.
8 Telfer P, Constantinidou G, Andreou P, *et al*. Quality of life in thalassemia. *Ann N Y Acad Sci* 2005;1054:273–82.
9 Lyrakos GN, Vini D, Aslani H, *et al*. Psychometric properties of the Specific Thalassemia Quality of Life Instrument for adults. *Patient Prefer Adherence* 2012;6:477–97.
10 Klaassen RJ, Barrowman N, Merelles-Pulcini M, *et al*. Validation and reliability of a disease-specific quality of life measure (the TranQol) in adults and children with thalassaemia major. *Br J Haematol* 2014;164:431–7.
11 Musallam KM, Khoury B, Abi-Habib R, *et al*. Health-related quality of life in adults with transfusion-independent thalassaemia intermedia compared to regularly transfused thalassaemia major: new insights. *Eur J Haematol* 2011;87:73–9.
12 Khoury B, Musallam KM, Abi-Habib R, *et al*. Prevalence of depression and anxiety in adult patients with β-thalassemia major and intermedia. *Int J Psychiatry Med* 2012;44:291–303.
13 Taher A, Viprakasit V, Cappellini MD, *et al*. Development of a patient-reported outcomes symptom measure for patients with nontransfusion-dependent thalassemia (NTDT-PRO©). *Am J Hematol* 2019;94:171–6.
14 FDA. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. 2009; https://www.fda.gov/media/77832/download (accessed 16 June 2022).
15 Taher A, Cappellini MD, Viprakasit V, *et al*. Validation of a patient-reported outcomes symptom measure for patients with nontransfusion-dependent thalassemia (NTDT-PRO©). *Am J Hematol* 2019;94:177–83.
16. Taher AT, Cappellini MD, Kattamis A, *et al*. Luspatercept for the treatment of anaemia in non-transfusion-dependent β-thalassaemia (BEYOND): a phase 2, randomised, double-blind, multicentre, placebo-controlled trial. *Lancet Haematol* 2022;9:e733–44.
17. Maruish ME. *User's Manual for the SF-36v2 Health Survey.* 3rd ed. Lincoln, RI: QualityMetric; 2011.
18. FACIT Group. FACIT-F scoring guidelines (version 4). 2021; www.facit.org/measures-scoring-downloads/facit-f-scoring-downloads (accessed 28 September 2021).
19. Hinkle DE, Wiersma W, Jurs SG. *Applied Statistics for the Behavioral Sciences*. 5th ed. Boston, MA: Houghton Mifflin; 2003.
20. Brown MB, Forsythe AB. Robust tests for the equality of variances. *J Am Stat Assoc* 1974;69:364–7.
21. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297–334.
22. Aaronson N, Alonso J, Burnam A, *et al*. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res* 2002;11:193–205.

23. Qin S, Nelson L, McLeod, L, *et al*. Assessing test-retest reliability of patient-reported outcome measures using intraclass correlation coefficients: recommendations for selecting and documenting the analytical formula. *Qual Life Res* 2019:28:1029–33.
24. Cappelleri JC, Zou KH, Bushmakin AG, *et al*. *Patient-Reported Outcomes: Measurement, Implementation and Interpretation.* Boca Raton, FL: CRC Press/Taylor & Francis; 2014.
25. Cella D, Lai JS, Chang CH, *et al*. Fatigue in cancer patients compared with fatigue in the general United States population. *Cancer* 2002;94:528–38.
26. Cella D, Eton DT, Lai JS, *et al*. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. *J Pain Symptom Manage* 2002;24:547–61.
27. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J Clin Epidemiol* 1997;50:869–79.
28. Nixon A, Doll H, Kerr C, *et al*. Interpreting change from patient reported outcome (PRO) endpoints: patient global ratings of concept versus patient global ratings of change, a case study among osteoporosis patients. *Health Qual Life Outcomes* 2016;14:25.

SUPPLEMENTARY MATERIALS

TABLE S1 ALGORITHM FOR MAPPING PRO ASSESSMENTS TO NOMINAL WEEKS

	Nominal week	NTDT-PRO	FACIT-F/SF-36v2®/PGI-C
Baseline	0	Days –7 to –1	Day of dosing of the first dose of study drug (screening if missing)
Weeks 1–12	1	Days 1 to 7	None
	2	Days 8 to 14	None
	3	Days 15 to 21	None
	4	Days 22 to 28	None
	5	Days 29 to 35	None
	6	Days 36 to 42	Days 22 to 63
	7	Days 43 to 49	None
	8	Days 50 to 56	None
	9	Days 57 to 63	None
	10	Days 64 to 70	None
	11	Days 71 to 77	None
	12	Days 78 to 84	Days 64 to 105
Weeks 13–24	13	Days 85 to 91	None
	14	Days 92 to 98	None
	15	Days 99 to 105	None
	16	Days 106 to 112	None
	17	Days 113 to 119	None
	18	Days 120 to 126	Days 106 to 147
	19	Days 127 to 133	None
	20	Days 134 to 140	None
	21	Days 141 to 147	None
	22	Days 148 to 154	None
	23	Days 155 to 161	None
	24	Days 162 to 168	Days 148 to 189

FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; NTDT, non-transfusion-dependent β-thalassaemia; PRO, patient-reported outcomes; SF-36v2®, Short Form Health Survey version 2.

Table S2 Known-groups validity at week 24

	n	NTDT-PRO T/W domain			NTDT-PRO SoB domain		
		LS mean	95% CI	*p* value[a]	LS mean	95% CI	*p* value[a]
FACIT-F FS				<0.001			<0.001
Very severe (≤37)	62	4.04	3.39, 4.69		3.67	2.99, 4.36	
Severe (>37 to 40)	16	2.63	1.61, 3.65		2.14	1.06, 3.22	
Moderate (>40 to 43)	18	2.52	1.59, 3.45		2.50	1.52, 3.48	
Mild (>43 to 46)	17	2.31	1.40, 3.23		2.01	1.04, 2.98	
Very mild/no symptoms (>46)	31	1.05	0.27, 1.82		0.62	–0.21, 1.44	
FACIT-F item HI12[b]				<0.001			<0.001
Very much (0)	3	6.57	4.68, 8.46		4.93	2.79, 7.07	
Quite a bit (1)	10	4.44	3.39, 5.49		3.85	2.67, 5.04	
Somewhat (2)	16	3.29	2.45, 4.12		3.39	2.44, 4.33	
A little bit (3)	40	2.77	2.20, 3.34		2.36	1.72, 3.00	
Not at all (4)	27	1.23	0.55, 1.92		0.93	0.16, 1.71	
FACIT-F item An2[b]				<0.001			0.002
Very much (0)	3	6.62	4.57, 8.68		4.92	2.68, 7.17	
Quite a bit (1)	11	4.08	3.01, 5.16		3.41	2.23, 4.58	
Somewhat (2)	15	3.36	2.45, 4.27		3.59	2.59, 4.58	
A little bit (3)	48	2.34	1.76, 2.93		1.96	1.32, 2.60	
Not at all (4)	19	1.78	0.91, 2.65		1.31	0.36, 2.26	
SF-36v2® vitality				<0.001			<0.001
Very poor (≤36.6)	7	5.37	4.07, 6.67		4.53	3.10, 5.96	
Poor (>36.6 to 43.3)	11	4.45	3.41, 5.49		4.04	2.90, 5.18	
Normal (>43.3 to 56.7)	41	2.98	2.40, 3.56		2.79	2.15, 3.43	
Better (>56.7 to 63.4)	29	1.72	1.05, 2.39		1.25	0.51, 1.98	
Much better (>63.4)	8	1.56	0.31, 2.80		1.48	0.11, 2.84	
SF-36v2® item 9e[c]				<0.001			0.001
All of the time (1)	3	3.13	1.10, 5.17		1.55	–0.72, 3.82	
Most of the time (2)	40	1.79	1.20, 2.39		1.58	0.92, 2.25	
Some of the time (3)	30	2.99	2.34, 3.64		2.76	2.03, 3.48	
A little of the time (4)	15	4.06	3.12, 5.00		3.51	2.47, 4.56	
None of the time (5)	8	5.13	3.88, 6.39		4.44	3.04, 5.85	
SF-36v2® item 9g[c]				<0.001			<0.001
All of the time (1)	5	5.67	4.24, 7.09		4.67	3.11, 6.24	
Most of the time (2)	4	5.03	3.35, 6.71		4.58	2.74, 6.43	
Some of the time (3)	18	3.79	3.01, 4.58		3.57	2.71, 4.43	
A little of the time (4)	44	2.62	2.07, 3.16		2.37	1.77, 2.97	
None of the time (5)	25	1.20	0.51, 1.90		0.78	0.02, 1.54	
SF-36v2® item 9i[c]				<0.001			<0.001
All of the time (1)	3	6.20	4.23, 8.17		6.47	4.30, 8.64	
Most of the time (2)	17	4.36	3.53, 5.19		3.56	2.64, 4.47	
Some of the time (3)	25	2.77	2.03, 3.50		2.53	1.72, 3.34	
A little of the time (4)	44	1.99	1.42, 2.56		1.76	1.14, 2.39	
None of the time (5)	7	1.58	0.25, 2.91		1.49	0.02, 2.96	

	n	NTDT-PRO T/W domain			NTDT-PRO SoB domain		
		LS mean	95% CI	*p* value[a]	LS mean	95% CI	*p* value[a]
PGI-S				<0.001			<0.001
0 to 2 (no symptoms)	43	1.13	0.72, 1.54		0.93	0.37, 1.48	
>2 to 4 (mild)	33	3.43	2.97, 3.89		3.32	2.69, 3.94	
>4 to 6 (moderate)	21	4.31	3.70, 4.91		3.63	2.82, 4.44	
>6 to 8 (severe)	11	5.60	4.85, 6.34		4.99	3.99, 6.00	
>8 (very severe)	2	6.81	5.07, 8.55		4.34	1.99, 6.69	

[a]*F*-test comparing T/W and SoB domain scores across subgroups (ANCOVA).

[b]"Please select 1 answer […] to indicate your response as it applies to the past 7 days": item HI12, "I feel weak all over"; item An2, "I feel tired".

[c]"How much of the time during the past week did you…": item 9e, "…have a lot of energy?"; item 9g, "…feel worn out?"; item 9i, "…feel tired?"

ANCOVA, analysis of covariance; CI, confidence interval; FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FS, Fatigue Subscale; LS, least squares; PGI-S, Patient Global Impression of Severity; SF-36v2®, Short Form Health Survey version 2; SoB, Shortness of Breath; T/W, Tiredness/Weakness.

Table S3 Demographics and baseline clinical characteristics

Characteristic	N=145
Age (years)	
Mean (SD)	39.9 (12.8)
Median (range)	40 (18 to 71)
Female, n (%)	82 (56.6)
Race, n (%)	
Asian	44 (30.3)
White	87 (60.0)
Other	14 (9.7)
Ethnicity, n (%)	
Hispanic or Latino	3 (2.1)
Not Hispanic or Latino	142 (97.9)
Body mass index (kg/m^2), n (%)	
<20	53 (36.6)
20 to <25	66 (45.5)
25 to <30	21 (14.5)
≥30	5 (3.5)
Geographic region, n (%)	
North America and Europe	90 (62.1)
Middle East	17 (11.7)
Asia Pacific	38 (26.2)
β-thalassaemia diagnosis, n (%)	
β-thalassaemia	97 (66.9)
Haemoglobin E/β-thalassaemia	39 (26.9)
β-thalassaemia plus α-thalassaemia	9 (6.2)
Baseline haemoglobin level (g/dL)	
Mean (SD)	8.2 (1.2)
Median (range)	8.2 (7.3 to 9.2)
Categories of baseline haemoglobin level, n (%)	
≥8.5 g/dL	60 (41.4)
<8.5 g/dL	85 (58.6)
Baseline transfusion burden (units of red blood cells in the 24 weeks before the first dose of study drug)	
Mean (SD)	0.3 (0.9)
Median (range)	0 (0 to 6)
6-minute walk test, n (%)	
≤450 m	82 (56.6)
>450 m	63 (43.4)
Left ventricular ejection fraction (%)	
Mean (SD)	65.6 (5.5)
Median (range)	65.0 (55.4 to 79.0)
Tricuspid valve regurgitation velocity, n (%)	
≤2.8 m/s (low probability of pulmonary hypertension)	111 (76.6)
>3.4 m/s (high probability of pulmonary hypertension)	1 (0.7)
ECOG performance status, n (%)	
0	100 (69.0)
1	45 (31.0)

ECOG, Eastern Cooperative Oncology Group; SD, standard deviation.

Table S4 Completeness of NTDT-PRO item entry at baseline and week 24

Number of days with missing NTDT-PRO data[a]	n (%)	
	Baseline (N=145)	Week 24 (N=128)
0	56 (38.6)	51 (39.8)
1	44 (30.3)	31 (24.2)
2	24 (16.6)	20 (15.6)
3	19 (13.1)	6 (4.7)
4	1 (0.7)	10 (7.8)
5	1 (0.7)	7 (5.5)
6	0	3 (2.3)
7	0	0

[a]There was no item-level missing data (participants either completed all 6 NTDT-PRO items or none of them).

NTDT-PRO, non-transfusion-dependent β-thalassaemia-patient-reported outcomes.

Table S5 Baseline PRO score distributions

	Mean (SD)	Median (Q1, Q3)	Range	Skewness	Kurtosis	Floor effect (%)[a]	Ceiling effect (%)[b]
NTDT-PRO							
Item 1-TiredNA	3.2 (2.2)	3.0 (1.5, 4.8)	0.0 to 9.0	0.2	–0.6	11.7	0.0
Item 2-TiredPA	5.0 (2.5)	5.2 (3.4, 7.0)	0.0 to 10.0	–0.3	–0.7	1.4	2.1
Item 3-WeakNA	3.1 (2.2)	3.0 (1.3, 4.8)	0.0 to 9.3	0.3	–0.5	11.7	0.7
Item 4-WeakPA	4.9 (2.6)	5.0 (3.0, 7.0)	0.0 to 10.0	–0.2	–0.8	2.8	2.1
Item 5-SobNA	2.4 (2.1)	2.2 (0.3, 4.0)	0.0 to 8.9	0.7	–0.2	20.7	0.0
Item 6-SobPA	4.2 (2.7)	4.4 (2.0, 6.4)	0.0 to 10.0	0.1	–1.0	7.6	2.8
T/W domain (items 1 to 4)	4.1 (2.2)	4.3 (2.5, 5.7)	0.0 to 9.5	0.0	–0.6	1.4	1.4
SoB domain (items 5 and 6)	3.3 (2.3)	3.4 (1.2, 5.1)	0.0 to 9.4	0.2	–0.8	7.6	0.7
PGI-S	3.7 (2.4)	3.8 (1.8, 5.4)	0.0 to 9.5	0.1	–0.8		
SF-36v2®							
Physical functioning	47.7 (7.7)	48.0 (44.2, 53.7)	23.1 to 57.5	–0.8	0.2	–	–
Role-physical	47.6 (7.8)	48.2 (41.4, 54.9)	25.7 to 57.2	–0.4	–0.7	–	–
Bodily pain	51.5 (9.2)	51.5 (42.6, 62.0)	30.6 to 62.0	–0.3	–1.1	–	–
General health	42.2 (10.2)	41.3 (34.2, 50.8)	19.0 to 66.5	0.1	–0.6	–	–
Vitality	49.2 (10.6)	49.6 (40.7, 58.5)	25.9 to 70.4	–0.3	–0.9	–	–
Social functioning	46.7 (9.3)	47.3 (37.3, 57.3)	22.3 to 57.3	–0.5	–0.8	–	–
Role-emotional	46.6 (8.8)	49.2 (38.8, 52.7)	17.9 to 56.2	–0.7	–0.4	–	–
Mental health	47.2 (9.6)	48.3 (40.4, 56.1)	24.7 to 64.0	–0.5	–0.6	–	–
PCS	48.0 (7.1)	48.8 (43.1, 53.3)	28.4 to 63.6	–0.4	–0.1	–	–
MCS	46.9 (9.2)	47.7 (40.6, 53.9)	23.3 to 63.1	–0.5	–0.4	–	–
FACIT-F							
Physical well-being	22.9 (3.9)	24.0 (20.0, 26.0)	11.0 to 28.0	–0.8	0.0	–	–
Social/family well-being	19.4 (5.3)	20.0 (16.3, 23.0)	4.7 to 28.0	–0.4	–0.5	–	–
Emotional well-being	18.2 (3.5)	19.0 (16.0, 21.0)	8.0 to 24.0	–0.6	–0.4	–	–
Functional well-being	18.0 (5.4)	18.0 (14.0, 22.0)	3.0 to 28.0	0.0	–0.6	–	–
FACT-G total score	78.4 (14.6)	80.0 (67.0, 90.3)	42.0 to 105.8	–0.1	–0.7	–	–
FACIT-F FS	36.4 (9.9)	39.0 (29.0, 44.5)	1.0 to 51.0	–0.7	0.0	–	–
FACIT-F TOI	77.2 (17.2)	81.0 (64.0, 91.0)	29.0 to 105.0	–0.4	–0.7	–	–
FACIT-F total score	114.8 (22.8)	118.5 (100.0, 133.2)	62.0 to 155.8	–0.3	–0.7	–	–

[a]Score of 0.

[b]Score of >9.

FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FACT-G, Functional Assessment of Cancer Therapy – General; FS, Fatigue Subscale; MCS, Mental Component Summary; PCS, Physical Component Summary; PGI-S, Patient Global Impression of Severity; PRO, patient-reported outcome; Q1, first quartile; Q3, third quartile; SD, standard deviation; SF-36v2®, Short Form Health Survey version 2; SoB, Shortness of Breath; SobNA, shortness of breath not doing physical activity; SobPA, shortness of breath doing physical activity; TiredNA, tiredness not doing physical activity; TiredPA, tiredness doing physical activity; TOI, trial outcome index; T/W, Tiredness/Weakness; WeakNA, weakness not doing physical activity; WeakPA, weakness doing physical activity.

Table S6 Variability of weekly NTDT-PRO item scores across missing day scenarios

		Number of missing days						
		0	**1**	**2**	**3**	**4**	**5**	**6**
Item 1-TiredNA	Mean	2.36	2.36	2.37	2.39	2.31	2.33	2.30
	SD	1.913	1.913	1.917	1.908	1.930	1.931	1.947
	p value[a]	–	0.971	0.949	0.971	0.962	0.869	0.962
Item 2-TiredPA	Mean	4.44	4.44	4.44	4.42	4.46	4.44	4.45
	SD	2.315	2.319	2.308	2.316	2.328	2.352	2.338
	p value[a]	–	1.000	0.953	0.970	0.978	0.827	0.873
Item 3-WeakNA	Mean	2.60	2.60	2.61	2.61	2.59	2.58	2.60
	SD	1.879	1.872	1.872	1.877	1.895	1.917	1.961
	p value[a]	–	0.941	0.930	0.955	0.888	0.786	0.576
Item 4-WeakPA	Mean	4.42	4.42	4.42	4.40	4.44	4.43	4.44
	SD	2.378	2.381	2.392	2.396	2.365	2.369	2.416
	p value[a]	–	0.997	0.973	0.892	0.871	0.965	0.764
Item 5-SobNA	Mean	2.02	2.02	2.01	2.03	2.01	2.05	2.05
	SD	1.894	1.892	1.884	1.911	1.884	1.939	1.928
	p value[a]	–	0.997	0.940	0.911	0.945	0.772	0.788
Item 6-SobPA	Mean	3.76	3.77	3.75	3.76	3.76	3.79	3.74
	SD	2.547	2.546	2.546	2.555	2.548	2.566	2.596
	p value[a]	–	0.982	0.970	0.958	0.993	0.859	0.849

The mean and SD were calculated by first calculating the average score across all weeks for each participant and then calculating the mean and SD across participants.

[a]Brown–Forsythe test comparing SD values for individual missing day scenarios with the SD when 0 days were missing.

NTDT-PRO, non-transfusion-dependent β-thalassaemia-patient-reported outcomes; SD, standard deviation; SobNA, shortness of breath not doing physical activity; SobPA, shortness of breath doing physical activity; TiredNA, tiredness not doing physical activity; TiredPA, tiredness doing physical activity; WeakNA, weakness not doing physical activity; WeakPA, weakness doing physical activity.

Table S7 NTDT-PRO internal consistency reliability

	Domain	Cronbach's alpha	Deleted item[a]	Cronbach's alpha
Baseline	T/W	0.95		
			Item 1-TiredNA	0.93
			Item 2-TiredPA	0.94
			Item 3-WeakNA	0.94
			Item 4-WeakPA	0.94
	SoB	0.89		
Week 24	T/W	0.94		
			Item 1-TiredNA	0.92
			Item 2-TiredPA	0.92
			Item 3-WeakNA	0.92
			Item 4-WeakPA	0.92
	SoB	0.85		
Weeks 13–24	T/W	0.95		
			Item 1-TiredNA	0.93
			Item 2-TiredPA	0.93
			Item 3-WeakNA	0.93
			Item 4-WeakPA	0.93
	SoB	0.84		

[a]The effect of removing individual items could not be evaluated for the SoB domain, because it consists of only 2 items.

NTDT-PRO, non-transfusion-dependent β-thalassaemia-patient-reported outcomes; SoB, Shortness of Breath; TiredNA, tiredness not doing physical activity; TiredPA, tiredness doing physical activity; WeakNA, weakness not doing physical activity; WeakPA, weakness doing physical activity; T/W, Tiredness/Weakness.

Table S8 Known-groups validity at baseline

		NTDT-PRO T/W domain			NTDT-PRO SoB domain		
	n	LS mean	95% CI	*p* value[a]	LS mean	95% CI	*p* value[a]
FACIT-F FS				<0.001			<0.001
Very severe (≤37)	62	5.27	4.84, 5.71		4.35	3.79, 4.91	
Severe (>37 to 40)	16	3.06	2.33, 3.80		3.30	2.36, 4.24	
Moderate (>40 to 43)	18	3.16	2.45, 3.86		2.84	1.93, 3.75	
Mild (>43 to 46)	17	2.94	2.21, 3.68		1.74	0.79, 2.68	
Very mild/no symptoms (>46)	31	1.59	1.05, 2.13		1.13	0.44, 1.83	
FACIT-F item HI12[b]				<0.001			<0.001
Very much (0)	3	7.11	5.47, 8.75		6.23	4.10, 8.36	
Quite a bit (1)	25	5.76	5.16, 6.35		4.80	4.03, 5.57	
Somewhat (2)	24	4.69	4.04, 5.34		4.06	3.22, 4.90	
A little bit (3)	54	3.58	3.18, 3.99		3.08	2.55, 3.60	
Not at all (4)	38	1.71	1.23, 2.18		1.15	0.54, 1.77	
FACIT-F item An2[b]				<0.001			<0.001
Very much (0)	3	7.87	6.21, 9.54		8.02	5.91, 10.13	
Quite a bit (1)	25	5.87	5.26, 6.48		4.89	4.11, 5.66	
Somewhat (2)	37	4.31	3.79, 4.83		3.90	3.24, 4.56	
A little bit (3)	59	3.08	2.68, 3.48		2.31	1.80, 2.82	
Not at all (4)	20	1.43	0.79, 2.08		1.26	0.44, 2.08	
SF-36v2® vitality				<0.001			<0.001
Very poor (≤36.6)	20	6.14	5.43, 6.84		5.57	4.66, 6.48	
Poor (>36.6 to 43.3)	19	5.42	4.70, 6.15		4.11	3.17, 5.05	
Normal (>43.3 to 56.7)	64	3.73	3.32, 4.13		3.15	2.63, 3.68	
Better (>56.7 to 63.4)	25	2.09	1.48, 2.69		1.73	0.95, 2.51	
Much better (>63.4)	13	1.71	0.90, 2.52		1.12	0.07, 2.17	
SF-36v2® item 9e[c]				<0.001			<0.001
All of the time (1)	11	2.09	1.14, 3.04		1.17	–0.02, 2.37	
Most of the time (2)	33	2.21	1.64, 2.77		1.95	1.24, 2.65	
Some of the time (3)	46	3.79	3.27, 4.31		3.24	2.59, 3.89	
A little of the time (4)	37	5.12	4.52, 5.73		4.18	3.42, 4.93	
None of the time (5)	14	5.80	4.91, 6.70		5.06	3.94, 6.19	
SF-36v2® item 9g[c]				<0.001			<0.001
All of the time (1)	1	6.64	3.62, 9.66		5.74	2.00, 9.47	
Most of the time (2)	24	5.67	5.02, 6.32		4.79	3.99, 5.59	
Some of the time (3)	39	4.43	3.92, 4.93		3.97	3.35, 4.60	
A little of the time (4)	41	2.78	2.27, 3.29		2.24	1.60, 2.87	
None of the time (5)	36	2.07	1.54, 2.60		1.40	0.75, 2.06	
SF-36v2® item 9i[c]				<0.001			<0.001
All of the time (1)	5	8.00	6.69, 9.31		7.70	6.01, 9.38	
Most of the time (2)	36	5.26	4.73, 5.79		4.34	3.66, 5.03	
Some of the time (3)	45	4.14	3.66, 4.61		3.58	2.97, 4.19	
A little of the time (4)	44	2.66	2.21, 3.11		2.08	1.50, 2.66	
None of the time (5)	11	1.21	0.35, 2.08		0.94	–0.18, 2.05	

		NTDT-PRO T/W domain			NTDT-PRO SoB domain		
	n	**LS mean**	**95% CI**	***p* value[a]**	**LS mean**	**95% CI**	***p* value[a]**
PGI-S				<0.001			<0.001
0 to 2 (no symptoms)	40	1.33	0.95, 1.71		1.06	0.51, 1.60	
>2 to 4 (mild)	37	3.70	3.31, 4.10		2.83	2.27, 3.40	
>4 to 6 (moderate)	44	4.90	4.52, 5.29		4.08	3.53, 4.63	
>6 to 8 (severe)	19	5.75	5.21, 6.30		5.17	4.39, 5.96	
>8 (very severe)	5	7.70	6.67, 8.72		7.43	5.96, 8.91	

[a]*F*-test comparing T/W and SoB domain scores across subgroups (ANCOVA).

[b]"Please select 1 answer […] to indicate your response as it applies to the past 7 days": item HI12, "I feel weak all over"; item An2, "I feel tired".

[c]"How much of the time during the past week did you…": item 9e, "…have a lot of energy?"; item 9g, "…feel worn out?"; item 9i, "…feel tired?"

ANCOVA, analysis of covariance; CI, confidence interval; FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FS, Fatigue Subscale; LS, least squares; NTDT-PRO, non-transfusion-dependent β-thalassaemia-patient-reported outcomes; PGI-S, Patient Global Impression of Severity; SF-36v2®, Short Form Health Survey version 2; SoB, Shortness of Breath; T/W, Tiredness/Weakness.

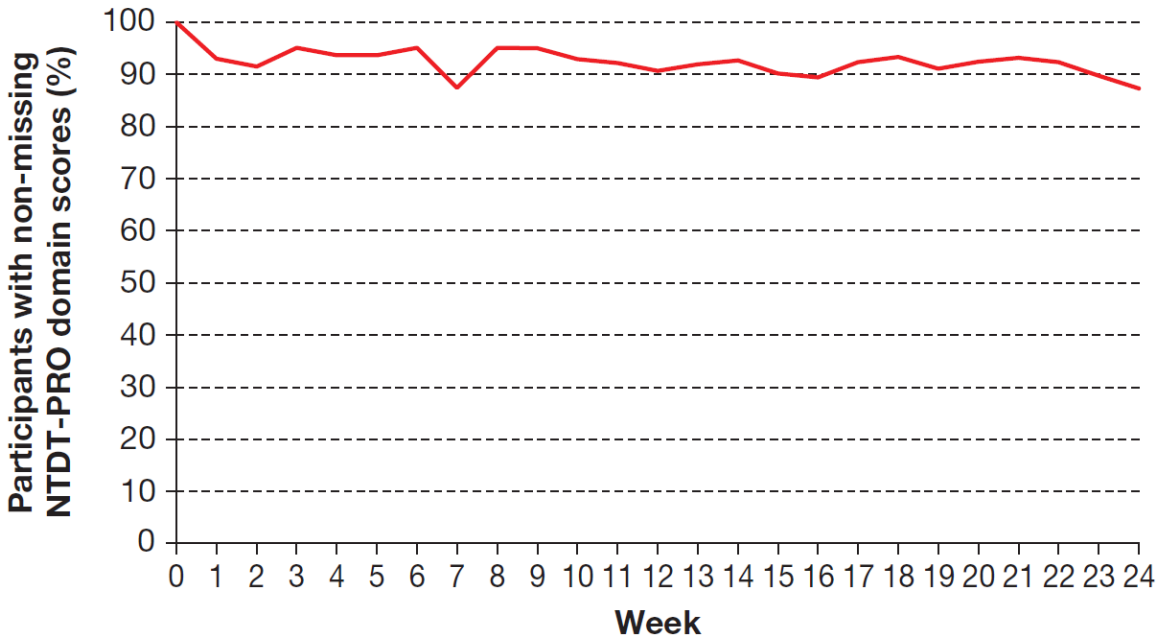

Figure S1 Percentage of participants with non-missing weekly NTDT-PRO domain scores**.** The percentage for a given week was calculated as the number of participants with non-missing weekly NTDT-PRO domain scores divided by the number of participants who remained on-study. For all weeks, percentages were the same for both the T/W and SoB domains.
NTDT-PRO, non-transfusion-dependent β-thalassaemia-patient-reported outcomes; SoB, Shortness of Breath; T/W, Tiredness/Weakness.

STROBE Statement—Checklist of items that should be included in reports of ***cross-sectional studies***

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract **Page 3, Abstract**
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found **Page 3, Abstract**
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported **Pages 6 and 7**
Objectives	3	State specific objectives, including any prespecified hypotheses **Page 7**
Methods		
Study design	4	Present key elements of study design early in the paper **Page 7, Study Design**
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection **Page 7, Study Design**
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants **Page 8, Participants**
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable **Pages 8 to 11, PRO assessments**
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group **Pages 8 to 11, PRO assessments**
Bias	9	Describe any efforts to address potential sources of bias **Not applicable**
Study size	10	Explain how the study size was arrived at **Page 17, Quality of completion of the NTDT-PRO**
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why **Pages 11 to 14, Statistical analyses**
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding **Page 11 to 14, Statistical analyses**
		(*b*) Describe any methods used to examine subgroups and interactions **Page 21, Known-groups validity**
		(*c*) Explain how missing data were addressed **Page 8 to 9, PRO assessments; Pages 11 and 12, Statistical analyses**
		(*d*) If applicable, describe analytical methods taking account of sampling strategy **Not applicable**
		(*e*) Describe any sensitivity analyses **Not applicable**
Results		

Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed **Page 17, Participants, Quality of completion of the NTDT-PRO**
		(b) Give reasons for non-participation at each stage **Not applicable**
		(c) Consider use of a flow diagram **Not applicable**
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders **Page 17, Participants**
		(b) Indicate number of participants with missing data for each variable of interest **Page 17, Quality of completion of NTDR-PRO**
Outcome data	15*	Report numbers of outcome events or summary measures **Not applicable**
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included **Tables 1 to 4**
		(*b*) Report category boundaries when continuous variables were categorized **Table 4**
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period **Not applicable**
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses **Not applicable**
Discussion		
Key results	18	Summarise key results with reference to study objectives **Page 24**
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias **Page 25**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence **Pages 24 to 25**
Generalisability	21	Discuss the generalisability (external validity) of the study results **Page 25**
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based **Page 26**

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely

available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.