BMJ Open

# 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]

**To cite:** Taher AT, Musallam KM, Viprakasit V, *et al*. Psychometric evaluation of the NTDT-PRO questionnaire for assessing symptoms in patients with non-transfusion-dependent beta-thalassaemia. *BMJ Open* 2023;**13**:e066683. doi:10.1136/bmjopen-2022-066683

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-066683).

Received 15 July 2022
Accepted 05 March 2023

For numbered affiliations see end of article.

**Correspondence to**
Dr Ali T Taher;
ataher@aub.edu.lb

## ABSTRACT

**Objectives** The non-transfusion-dependent beta-thalassaemia-patient-reported outcome (NTDT-PRO) questionnaire was developed for assessing anaemia-related tiredness/weakness (T/W) and shortness of breath (SoB) among patients with 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 ≤100 g/L.

**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 PROs 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 patient-reported outcome (PRO) instruments such as Patient Global Impression of Severity, Patient Global Impression of Change, 36-Item Short Form Health Survey version 2 and Functional Assessment of Chronic Illness Therapy–Fatigue.

⇒ The data used in this analysis were from a phase 2 interventional study with participants from multiple geographical regions and spanning a range of non-transfusion-dependent beta-thalassaemia (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

Beta-thalassaemias are a group of genetic blood disorders characterised by defective synthesis of the beta-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 beta-thalassaemia (TDT) or non-transfusion-dependent beta-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 accumulates as patients get older, NTDT can result in various comorbidities (eg, hepatic disease, endocrinopathy, thromboembolic events, pulmonary hypertension, leg ulcers and extramedullary haematopoietic 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 beta-thalassaemias has typically been evaluated by generic questionnaires such as the Short Form Health Survey version 2 (SF-36v2) and the WHO 100-item Quality of Life Survey,[7 8] which may fail to capture the unique experiences of patients with beta-thalassaemia. Two beta-thalassaemia-specific PRO instruments for assessing HRQoL are now available: the Specific Thalassaemia Quality of Life Instrument and the Transfusion-dependent Quality of Life 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 beta-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 six-item questionnaire intended to measure the most relevant and important anaemia-related symptoms of NTDT.[13] In accordance with US Food and Drug Administration 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.

The psychometric analysis plan was finalised prior to the finalisation of the core study statistical analysis plan and study unblinding. All analyses were carried out on an interim blinded data cut, and all analysts remained blinded until programming of all prespecified analyses was complete.

### Participants

Participants were adults (≥18 years of age) with beta-thalassaemia or haemoglobin E/beta-thalassaemia. They were non-transfusion dependent, as defined by receipt of 0–5 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 two measurements taken ≥1 week apart) of ≤100 g/L and an Eastern Cooperative Oncology Group (ECOG) performance status of 0 or 1. Patients with haemoglobin S/beta-thalassaemia or alpha-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 geographical 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 supplemental table 1).

### NTDT-PRO questionnaire

The NTDT-PRO assesses the severity of symptoms associated with NTDT in the 24 hours prior to administration. The six items assess tiredness (lack of energy, two items), weakness (lack of strength, two items) and shortness of breath (SoB) (two items) when doing and when not doing physical activity. Each item uses an 11-point Numerical 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 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 (ie, 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 two of the four T/W items (including ≥1 tiredness item and ≥1 weakness item) or at least one of the two SoB 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'.

### Patient Global Impression of Severity

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.

### Patient Global Impression of Change

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 one of seven response options ranging from 'a great deal better' to 'a great deal worse'.

### The 36-Item Short Form Health Survey version 2

SF-36v2 consists of eight 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 V.5 (QualityMetric, Lincoln, Rhode Island, 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 0–100 scale. They were then transformed to a US norm-based T-score (mean: 50, SD: 10), with a higher T-score indicating better health. Finally, the Physical Component Summary and Mental Component Summary (MCS) were derived as weighted averages of the T-scores for the eight multi-item scales.

### Functional Assessment of Chronic Illness Therapy–Fatigue

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

Scores for three additional summary scales were also calculated: FACT-G total score=sum of scores for all FACT-G items (range: 0–108 points); FACIT-F trial outcome index=sum of the scores for FACT-G physical well-being, FACT-G functional well-being and the FS (range: 0–108 points); and FACIT-F total score=sum of scores for all FACT-G items and the FS (range: 0–160 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 V.9.4 (SAS Institute, Cary, North Carolina, USA). Analyses were performed on blinded data collected up to week 24 during double-blind treatment (data cut-off: 7 January 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–<0.7=moderate, ≥0.7–<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 four T/W items into one domain and the two SoB 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 two-way mixed-effects analysis of variance 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 validity and discriminant validity were 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 supplemental table 2. 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 geographical 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 five 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 geographical region.

## RESULTS

### Participants

The ITT population comprised 145 participants with a mean (SD) age of 39.9 (12.8) years (range: 18–71 years) (see online supplemental table 3). 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/beta-thalassaemia, and 6.2% had a diagnosis of beta-thalassaemia combined with alpha-thalassaemia. The mean (SD) haemoglobin level at baseline was 82 (12) g/L, 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 (ie, with sufficient data to calculate average weekly item scores) was 98.6% at baseline and 84.4% at week 24 (see online supplemental table 4). 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 supplemental figure 1).

### 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 supplemental table 5). 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 supplemental table 5). 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 was no problematic floor or ceiling effects.

### NTDT-PRO item–item and item–domain correlations

Across the three 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=0.97–0.98), and item 2-TiredPA was very strongly correlated with item 4-WeakPA (weakness doing physical activity) (r=0.98–0.99). Item 5-SobNA and item 6-SobPA (shortness of breath doing physical activity) were strongly correlated with each other (r=0.74–0.81) and moderately to strongly correlated with item 1-TiredNA, item 2-TiredPA, item 3-WeakNA and item 4-WeakPA (r=0.50–0.81) (table 2).

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

### 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 supplemental table 6). Moreover, when comparing SD values for the different missing day scenarios using the Brown–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 0.94–0.95 across the three assessment time points/time intervals (baseline, week 24, weeks 13–24) (see online supplemental table 7), 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 0.84–0.89, also indicating acceptable internal consistency reliability.

Table 1 Responsiveness at weeks 13–24

| | Spearman's rank correlation coefficient (r)* | | LS mean change (95% CI) at weeks 13–24† | | | | |
|---|---|---|---|---|---|---|---|
| | Week 24 | Weeks 13–24 | Improvement level 2 | Improvement level 1 | No change | Worsening | P value‡ |
| **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 | – | – | – | – | – |

*Changes from baseline.

†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.

‡F-test comparing T/W and SoB domain scores across response categories (ANCOVA).

ANCOVA, analysis of covariance; FACIT-F, Functional Assessment of Chronic Illness Therapy–Fatigue; FS, Fatigue Subscale; LS, least squares; N/A, not applicable; NTDT-PRO, non-transfusion-dependent beta-thalassaemia-patient-reported outcome; 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.

**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 |
| 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 beta-thalassaemia-patient-reported outcome; 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; T/W, tiredness/weakness; WeakNA, weakness not doing physical activity; WeakPA, weakness doing physical activity.

**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* | | | | | | |
| 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† | | | | | | |
| 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‡ | 0.86 | 0.83 | 0.80 | 0.72 | 0.67 | 0.65 |

*n=141 at baseline, n=96 at week 24, n=125 at weeks 13–24.
†n=144 at baseline, n=96 at week 24, n=126 at weeks 13–24.
‡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 beta-thalassaemia-patient-reported outcome; 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.

### 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.

### 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 supplemental table 8) and week 24 (see online supplemental table 2), 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

**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*** | **LS mean** | **95% CI** | **P value*** |
| FACIT-F FS | | | | <0.001 | | | <0.001 |
| Very severe (≤37) | 43 | 4.39 | 3.90, 4.88 | | 3.90 | 3.35, 4.45 | |
| Severe (>37–40) | 16 | 2.91 | 2.10, 3.73 | | 1.77 | 0.86, 2.68 | |
| Moderate (>40–43) | 19 | 2.81 | 2.06, 3.55 | | 2.61 | 1.77, 3.45 | |
| Mild (>43–46) | 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† | | | | <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† | | | | <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 (>36.6–43.3) | 19 | 4.51 | 3.54, 5.48 | | 3.83 | 2.76, 4.89 | |
| Normal (>43.3–56.7) | 64 | 3.05 | 2.55, 3.55 | | 2.82 | 2.27, 3.37 | |
| Better (>56.7–63.4) | 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‡ | | | | <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‡ | | | | <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‡ | | | | <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 | |

Continued

Table 4 Continued

| | n | NTDT-PRO T/W domain | | | NTDT-PRO SoB domain | | |
|---|---|---|---|---|---|---|---|
| | | LS mean | 95% CI | P value* | LS mean | 95% CI | P value* |
| PGI-S | | | | <0.001 | | | <0.001 |
| 0–2 (no symptoms) | 45 | 1.37 | 0.94, 1.79 | | 1.10 | 0.57, 1.62 | |
| >2–4 (mild) | 36 | 2.93 | 2.47, 3.40 | | 2.68 | 2.10, 3.26 | |
| >4–6 (moderate) | 34 | 4.48 | 3.99, 4.98 | | 3.95 | 3.32, 4.57 | |
| >6–8 (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 | |

*F-test comparing T/W and SoB domain scores across subgroups (ANCOVA).
†'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'.
‡'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; FACIT-F, Functional Assessment of Chronic Illness Therapy–Fatigue; FS, Fatigue Subscale; LS, least squares; NTDT-PRO, non-transfusion-dependent beta-thalassaemia-patient-reported outcome; PGI-S, Patient Global Impression of Severity; SF-36v2, Short Form Health Survey version 2; SoB, shortness of breath; T/W, tiredness/weakness.

item 9e (having a lot of energy) and the PGI-S. Known-groups validity of the T/W and SoB domains was therefore demonstrated.

### 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 five 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 was as we hypothesised, and is 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 with 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 geographical 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 T/W and SoB 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.

**Author affiliations**
[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, New Jersey, USA
[7]Evidera, Waltham, Massachusetts, USA
[8]Evidera, Bogotá, Colombia
[9]Adelphi Values, Boston, Massachusetts, 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

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

**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 analyses. 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.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not required.

**Ethics approval** 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.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request. All data relevant to the study are included in the article or uploaded as supplemental information.

**ORCID iDs**
Ali T Taher http://orcid.org/0000-0001-8515-2238
Khaled M Musallam http://orcid.org/0000-0003-3935-903X
Christopher Pelligra http://orcid.org/0000-0002-5255-2777

## REFERENCES

1 Taher AT, Musallam KM, Cappellini MD. β-thalassemias. *N Engl J Med* 2021;384:727–43.
2 Musallam KM, Rivella S, Vichinsky E, *et al*. Non-transfusion-dependent thalassemias. *Haematologica* 2013;98:833–44.
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-naïve 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. Available: www.fda.gov/media/77832/download [Accessed 16 Jun 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(10):e733–e744.

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. Available: www.facit.org/measures-scoring-downloads/facit-f-scoring-downloads [Accessed 28 Sep 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. *Patient-Reported Outcomes: Measurement, Implementation and Interpretation*. Boca Raton, FL: CRC Press/Taylor & Francis, 2014.

25 Cella D, Lai J-S, Chang C-H, *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 J-S, *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.