



BMJ Open Identifying thresholds for meaningful improvements in NTDT-PRO scores to support conclusions about treatment benefit in clinical studies of patients with non-transfusion-dependent beta-thalassaemia: analysis of pooled data from a phase 2, double-blind, placebo-controlled, randomised trial

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 G Pelligra ⁸, Alan L Shields,⁹ Jeevan K Shetty,¹⁰ Mrudula B Glassberg,¹¹ Luciana Moro Bueno,¹⁰ Maria Domenica Cappellini¹²

To cite: Taher AT, Musallam KM, Viprakasit V, *et al*. Identifying thresholds for meaningful improvements in NTDT-PRO scores to support conclusions about treatment benefit in clinical studies of patients with non-transfusion-dependent beta-thalassaemia: analysis of pooled data from a phase 2, double-blind, placebo-controlled, randomised trial. *BMJ Open* 2024;**14**:e085234. doi:10.1136/bmjopen-2024-085234

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<https://doi.org/10.1136/bmjopen-2024-085234>).

Received 09 February 2024
Accepted 30 September 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Ali T Taher;
ataher@aub.edu.lb

ABSTRACT

Objectives To estimate thresholds for defining meaningful within-patient improvement from baseline to weeks 13–24 and interpreting meaningfulness of between-group difference for the non-transfusion-dependent beta-thalassaemia patient-reported outcome (NTDT-PRO) tiredness/weakness (T/W) and shortness of breath (SoB) scores. A secondary objective was to determine the symptom severity threshold for the NTDT-PRO T/W domain to identify patients with symptomatic T/W.

Design Pooled blinded data from the phase 2, double-blind, placebo-controlled, randomised BEYOND trial in NTDT (NCT03342404) were used. Anchor-based analyses supplemented with distribution-based analyses and empirical cumulative distribution function (eCDF) curves were applied. Distribution-based analyses and receiver operating characteristic curves were used to estimate between-group difference and symptomatic thresholds, respectively.

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

Participants Adults (N=145; mean age 39.9 years) with NTDT who were transfusion-free ≥8 weeks before randomisation.

Measures Score changes from baseline to weeks 13–24 in PROs used as anchors (correlation coefficient ≥0.3): NTDT-PRO T/W and SoB scores, Patient Global Impression of Severity, Functional Assessment of Chronic Illness Therapy–Fatigue (Fatigue Subscale, item HI12 and item An2) and Short Form Health Survey version 2.

Results The eCDF curves support the use of estimates from the improvement by one level group for all anchors to determine the threshold(s) for meaningful within-patient improvement. Mean (median) changes from these groups

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ Clinically meaningful within-patient thresholds for improvement were estimated using anchor-based analyses employed per relevant regulatory guidelines and the use of multiple anchors, measuring similar concepts of interest from patients' perspectives.
- ⇒ Moreover, to increase robustness, multiple methods were used to derive the estimates of the possible thresholds for triangulation.
- ⇒ The use of non-verbal rating scales was a limitation of this study.
- ⇒ Another limitation was that the initial symptomatic tiredness/weakness threshold estimation was based on the Functional Assessment of Chronic Illness Therapy–Fatigue, Fatigue Subscale threshold differentiating anaemic patients with cancer from the general population.
- ⇒ Finally, while the study enrolled patients from several geographic regions who had a wide range of symptom severity, this clinical trial population may not be representative of the larger beta-thalassaemia patient population.

and estimates from distribution-based analyses suggest that a ≥1-point reduction in the NTDT-PRO T/W or SoB domains represents a clinically meaningful improvement. Meaningful between-group difference threshold ranges were 0.53–1.10 for the T/W domain and 0.65–1.15 for the SoB domain. The optimal symptomatic threshold for the T/W domain (by maximum Youden's index) was ≥3 points.

Conclusions The thresholds proposed may support the use of NTDT-PRO in assessing and interpreting treatment effects in clinical studies and identifying patients with NTDT in need of symptom relief.

INTRODUCTION

Beta-thalassaemias are hereditary blood disorders caused by defective synthesis of the beta globin chains of haemoglobin A, leading to anaemia.¹ Approximately 1.5% of the world population (up to 80 million individuals) are beta-thalassaemia carriers.¹ Beta-thalassaemias can be classified based on transfusion requirements: patients with transfusion-dependent beta-thalassaemia (TDT) require lifelong regular blood transfusions to survive, while those with non-TDT (NTDT) do not require regular transfusions but may need transfusions occasionally or in specific clinical settings.²⁻⁷

In NTDT, ineffective erythropoiesis and peripheral haemolysis result in anaemia-related symptoms^{8,9} and clinical complications that cause a substantial burden, affect survival and impair quality of life.^{8 10-14} Such complications include skeletal deformities, hepatosplenomegaly and extramedullary haematopoiesis, pulmonary hypertension, leg ulcers, hepatic and endocrine disorders, thromboembolic events and iron overload.^{4 5 8} Current treatments are limited to on-demand red blood cell transfusions, splenectomy, fetal haemoglobin induction and iron chelation therapy.² Thus, alternative treatment options to improve ineffective erythropoiesis and related symptoms in patients with NTDT are needed.¹⁵ Furthermore, instruments to evaluate the effects of these treatments as reported by patients are lacking.

To assess treatment effect on NTDT symptoms from the patient perspective in the BEYOND study,¹⁶ a fit-for-purpose patient-reported outcome (PRO) instrument, the NTDT-PRO, was developed following US Food and Drug Administration (FDA) guidance.¹⁷ The NTDT-PRO is a six-item questionnaire for assessing the severity of tiredness/weakness (T/W) and shortness of breath (SoB), two of the most frequently reported NTDT symptoms in concept elicitation interviews with patients.¹¹ Evidence supporting the NTDT-PRO's content validity (ie, it measures concepts relevant to the disease and important to patients) and psychometric performance (ie, it generates reliable scores on which valid conclusions can be based) in the target patient population of patients with NTDT is well established.^{11 18 19}

However, there are currently no guidelines for the interpretation of changes (ie, a within-patient evaluation) in the scores generated by the NTDT-PRO over time, or observed score differences (ie, a between-treatment group comparison), that are meaningful to patients. Without this information, clinical researchers are unable to identify patients whose NTDT symptoms may have improved in severity by a meaningful amount when assessing treatment effect or to interpret the meaningfulness of a difference in mean NTDT-PRO scores between treatment groups in a clinical trial. Therefore,

the primary objectives of the present analysis were to estimate the thresholds for defining meaningful within-patient change from baseline and to interpret the meaningfulness of between-group differences for NTDT-PRO T/W and SoB domains in patients with NTDT. As patients with NTDT may have mild or no symptoms, thresholds to identify patients who require symptomatic treatment are also needed. Therefore, the secondary objective was to determine symptom severity threshold(s) for the NTDT-PRO T/W domain.

METHODS

Study design and participants

This analysis was based on pooled, blinded data collected up to week 24 in the phase 2 BEYOND trial of luspatercept in adults with NTDT (NCT03342404).¹⁶ The study design has been described elsewhere.^{16 19} Briefly, eligible patients were randomised 2:1 using an interactive response technology system to receive luspatercept or placebo subcutaneously every 3 weeks for 48 weeks during the double-blind treatment phase. Patients were stratified based on baseline haemoglobin concentration (≥ 8.5 g/dL vs < 8.5 g/dL) and baseline NTDT-PRO T/W domain score (≥ 3 vs < 3). The psychometric analysis plan was completed before core study statistical analysis plan finalisation and prior to study unblinding. All analyses were conducted on an interim blinded data set and remained blinded until completion of all prespecified analyses' programming. Masking success was determined by unmasked monitors. All analyses were based on the intention-to-treat population. The primary efficacy endpoint was the proportion of patients who had a ≥ 1.0 g/dL increase in mean haemoglobin from baseline over a continuous 12-week interval, from week 13 to week 24, without transfusion. Mean change from baseline in NTDT-PRO T/W domain score over the same time interval was a key secondary efficacy endpoint in the sequential testing for statistical significance.¹⁶ Therefore, weeks 13-24 were chosen as the time interval of interest for this analysis, as it represents the expected time for a sustainable response.

Assessments

The six items of the NTDT-PRO comprise two domains: a T/W domain with four items assessing tiredness (lack of energy) and weakness (lack of strength) when doing and not doing physical activity; and an SoB domain with two items assessing SoB when doing and not doing physical activity.^{11 18} Each item is scored on an 11-point numerical rating scale ranging from 0 (no symptoms) to 10 (extreme symptoms) and has a recall period of 24 hours. Weekly average item scores from baseline to week 24 were calculated by taking the average daily score for each item in each week.

Several additional measures administered in BEYOND contributed to the present analysis. Patients completed the Patient Global Impression of Severity (PGI-S) daily in the evening from 1 week prior to randomisation to

week 24. They also completed the Short Form Health Survey version 2 (SF-36v2),²⁰ the Functional Assessment of Chronic Illness Therapy–Fatigue (FACIT-F)²¹ and the PGI of Change (PGI-C)²² at screening and on the day of dosing of every other dose of study drug (ie, every 6 weeks).

Additional details (including scoring, how the various scales complement each other in terms of the assessed outcomes and handling of missing data) on the PROs administered in BEYOND are provided in online supplemental material 1.

Statistical analyses

Statistical analyses were conducted by using SAS V.9.4 or higher (SAS Institute). Analyses were performed using blinded data for all randomised participants.

Clinically meaningful within-patient threshold for improvement

Consistent with FDA guidance, an anchor-based analysis was implemented as the primary approach to estimate clinically meaningful within-patient improvement in T/W and SoB scores.¹⁷ The anchor-based approach uses an external criterion to categorise patients into a priori-determined groups with different levels of self-reported treatment response (eg, improvement, no change, worsening). Appropriate anchors should be described as plainly understood, assessing similar concepts to the concept measured by the target assessment (the NTDT-PRO in this case) and having sufficient correlation with the target PRO measure (correlation coefficient ≥ 0.3).

The use of multiple anchors is recommended by the FDA.¹⁷ Thus, the following clinical and PRO measures that were used in BEYOND alongside the NTDT-PRO were evaluated for their suitability as anchors for this analysis: haemoglobin level, PGI-S, PGI-C, FACIT-F Fatigue Subscale (FS), FACIT-F item HI7 ('I feel fatigued the past 7 days'), FACIT-F item HI12 ('I feel weak all over the past 7 days'), FACIT-F item An2 ('I feel tired the past 7 days'), FACIT-F item An5 ('I have energy the past 7 days'), SF-36v2 vitality, SF-36v2 item 9e ('How much of the time during the past week did you have a lot of energy?'), SF-36v2 item 9g ('How much of the time during the past week did you feel worn out?') and SF-36v2 item 9i ('How much of the time during the past week did you feel tired?'). Haemoglobin level was chosen because it is a well-established clinical outcome in NTDT and was used to define the primary efficacy endpoint in BEYOND.^{2 16 23} PGI-S and PGI-C, which measure the severity of overall NTDT-related symptoms and change in the overall symptoms, respectively, are anchors recommended by the FDA.¹⁷ The FACIT-F FS and SF-36v2 vitality scores are PRO domain scores measuring concepts related to the NTDT-PRO T/W domain with previously established clinically meaningful within-patient change thresholds described below. Finally, FACIT-F items HI7, HI12, An2 and An5, and SF-36v2 items 9e, 9g and 9i were chosen as they are

single-item Verbal Rating Scales (VRSs), each measuring concepts like those targeted by the NTDT-PRO T/W domain and having response options that could be easily interpreted to indicate different levels of change.

Spearman's rank correlation coefficients between changes in T/W and SoB domain scores from baseline to weeks 13–24 and changes in the 12 potential anchors over the same period were calculated (except for PGI-C, which is already a measure of change from the start of the study, where the absolute score at weeks 13–24 was used in the correlation calculations). Five of the potential anchors with the highest correlation coefficients (and absolute value ≥ 0.3) with both NTDT-PRO T/W and SoB domains were chosen to be used in the anchor-based analyses.^{24 25} Patients were then categorised by level of response on each of the five chosen anchors, and descriptive statistics on the change in NTDT-PRO T/W and SoB scores and corresponding empirical cumulative distribution function (eCDF) and probability distribution function (PDF, using the kernel density estimator) curves were generated for each of the levels of response. Levels of response were defined (see online supplemental table S1) based on the clinically meaningful within-patient improvement threshold on the anchors (for continuous scales), and their meaningfulness was confirmed on inspection of the eCDF curves. Meaningful improvement on the PGI-S was defined as a decrease of 1 point, and 4-point and 6.7-point increases on the FACIT-F FS and SF-36v2 vitality domains were chosen to reflect meaningful improvements based on the findings by the instruments' developers.^{20 26} For each of the FACIT-F and SF-36v2 VRS items included as anchors, a 1-point change (ie, one level change on the VRS) was defined as a meaningful change (also confirmed on inspection of the eCDF curves).

Distribution-based estimates, suggested as a supportive approach by the FDA, were given by the SE of measurement (SEM, as estimated based on the method provided in online supplemental material 1) and half of the SD at baseline of the NTDT-PRO T/W and SoB scores.^{17 27–30}

Mean and median changes in the NTDT-PRO T/W and SoB domain scores, obtained from the a priori-determined anchor group with the level(s) of improvement deemed to be meaningful (which were guided by the eCDF and PDF curves), were considered in triangulation of the final clinically meaningful within-patient improvement thresholds. Estimates from the receiver operating characteristic (ROC) curve analyses and distribution-based analyses were considered supportive in determining the thresholds, or ranges of thresholds, for each NTDT-PRO domain.

Finally, to assess the appropriateness of the newly derived meaningful improvement thresholds, the percentages of patients who would be considered responders on the NTDT-PRO T/W and SoB domains when applying the thresholds were calculated among those patients who achieved an average ≥ 1.0 g/dL (and ≥ 1.5 g/dL) change from baseline to weeks 13–24.

Clinically meaningful between-group difference threshold

Clinically meaningful between-group difference thresholds were estimated from the distribution-based approach by calculating the SEM and 0.5 SD at baseline for both the T/W and SoB domains. SEM was estimated by the baseline SD of the NTD-T-PRO T/W and SoB scores multiplied by the square root of one minus the reliability coefficient for each corresponding domain (ie, the intraclass correlation coefficient (ICC)).^{28 31} The ICC of the weekly domain scores between baseline and week 1 among stable patients was calculated using the two-way mixed-effect analysis of variance model, with the week as a fixed effect. Stable patients were considered those who had the same PGI-S weekly scores at baseline and week 1.

Symptomatic threshold

To estimate the symptomatic threshold for the NTD-T-PRO T/W domain score, ROC analysis was performed. All potential NTD-T-PRO T/W score thresholds were assessed for their accuracy at classifying symptomatic and less/asymptomatic participants, as defined using FACIT-F FS (comprising 13 items specific to fatigue) and SF-36v2 vitality (comprising questions about patients' perception of their energy levels and tiredness). These scales were selected as their concepts overlap with those that the NTD-T-PRO T/W aims to capture (ie, T/W). Patients with FACIT-F FS score <43 or SF-36v2 vitality score <45 were defined as more symptomatic and those with FACIT-F FS score ≥ 43 or SF-36v2 vitality score ≥ 45 as less/asymptomatic.^{20 32}

ROC analyses were conducted using pooled assessments from baseline and weeks 6, 12, 18 and 24, with area under the curve (AUC) values of 0.5, 0.7 and 1.0 indicating no diagnostic ability (ie, similar to random guessing), good diagnostic ability and perfect diagnostic accuracy, respectively. Similar to the analysis to estimate the meaningful within-patient improvement threshold, the NTD-T-PRO T/W score that maximised Youden's index was identified as the optimal cut-off above which scores indicate symptomatic disease and only those cut-off values from the ROC analyses with AUC ≥ 0.70 , indicating good performance, were considered.^{33 34}

Patient and public involvement

Patients and the public were not involved in the design, conduct, reporting or dissemination plans of this research. However, all participating patients provided informed consent.

RESULTS

Participants

The analysis included 145 participants, whose demographic and baseline clinical characteristics are described in a previous psychometric evaluation of the NTD-T-PRO.¹⁹ Briefly, patients had a mean (SD) age of 39.9 (12.8) years with a mean (SD) haemoglobin level of 82 (12) g/L. Baseline mean (SD) NTD-T-PRO T/W

and SoB domain scores were 4.1 (2.2) and 3.3 (2.3), respectively. Patients had a mean (SD) PGI-S score of 3.7 (2.4) at baseline.

Clinically meaningful within-patient improvement threshold estimates

Score changes in the NTD-T-PRO T/W and SoB domains from baseline to weeks 13–24 were at least moderately correlated with score changes in almost all of the 12 potential anchors in expected directions (ie, negative for haemoglobin, FACIT-F FS, FACIT-F items, SF-36 vitality and items 9g and 9i, and positive for PGI-S, PGI-C and SF-36 item 9e) (online supplemental table S2). The anchors chosen were those with the highest longitudinal correlation coefficients (absolute values) with both the T/W and SoB domain scores: PGI-S, FACIT-F FS, FACIT-F items HI12 and An2, and SF-36v2 vitality. Among the five anchoring assessments, PGI-S had the highest correlations (absolute values of 0.79 and 0.69 for the T/W and SoB domain scores, respectively) (online supplemental table S2).

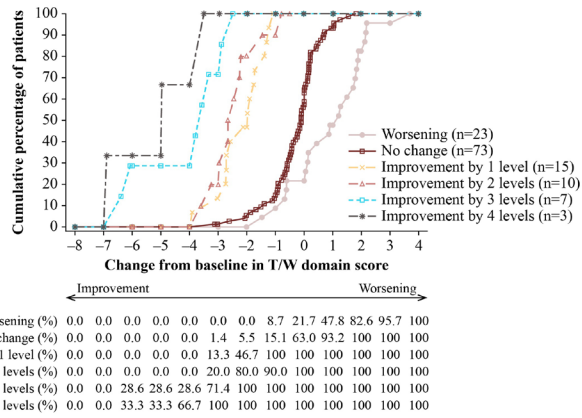
The eCDF curves showed clear separation between the improvement by one level and no change groups for all anchors for both the T/W (figure 1) and SoB domains (figure 2), suggesting that estimates from the groups with one level of improvement on these anchors can be considered to support the triangulation of meaningful within-patient improvement threshold(s) for the T/W and SoB domains. The corresponding PDF curves, providing an overview of the shape, dispersion and skewness of the distribution of score changes in the T/W and SoB domains for each of the anchor response groups, are shown in online supplemental figures S1 and S2.

T/W domain score

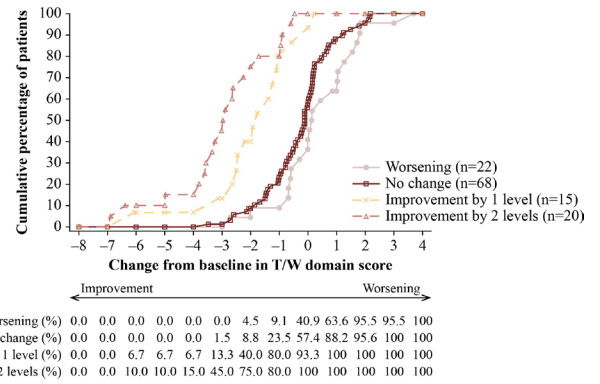
The direction and magnitude of mean and median changes in T/W score from baseline to weeks 13–24 were consistent with the levels of improvement on each anchor, with a larger decrease in T/W score (ie, improvement) associated with a higher degree of improvement on each anchor (table 1). Mean changes from baseline (effect size) in the T/W domain scores for the group with one level of improvement ranged from -1.31 (-0.59 , FACIT-F item An2) to -2.14 (-0.98 , PGI-S) and corresponded to moderate to large effect sizes (ie, ≥ 0.5 in absolute value).³⁵ Median changes ranged from -1.47 (FACIT-F item An2) to -2.06 (FACIT-F item HI12). Distribution-based analyses gave estimates of 1.10 (0.5 SD) and 0.53 (SEM), with estimates from the ROC analyses indicating optimal thresholds (maximising Youden's index) to be -0.91 (FACIT-F FS and FACIT-F item HI12), -1.05 (SF-36v2 vitality), -1.08 (PGI-S) and -1.47 (FACIT-F item An2). All ROC analyses exceeded the AUC threshold of 0.70, indicating good discriminant power.

Based on these findings, a ≥ 1.0 -point decrease in T/W score (on a scale of 0–10) was considered to represent a lower bound for the clinically meaningful

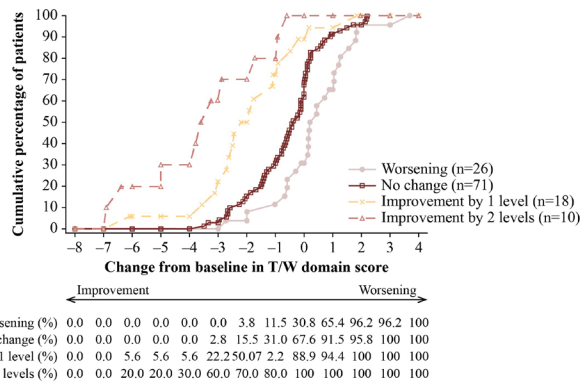
A PGI-S*



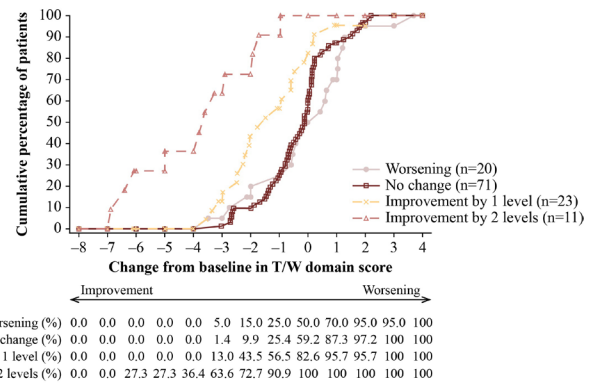
B FACIT-F FS†



C FACIT-F item HI12‡



D FACIT-F item An2‡



E SF-36v2 vitality§

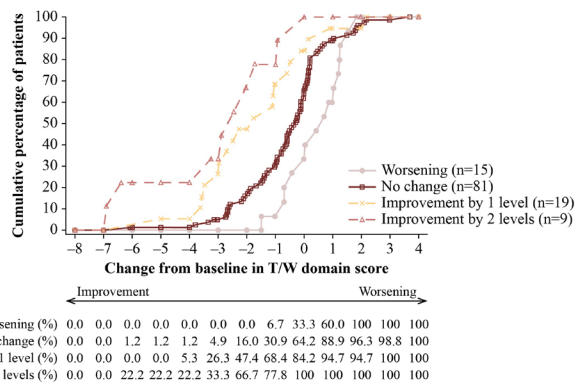
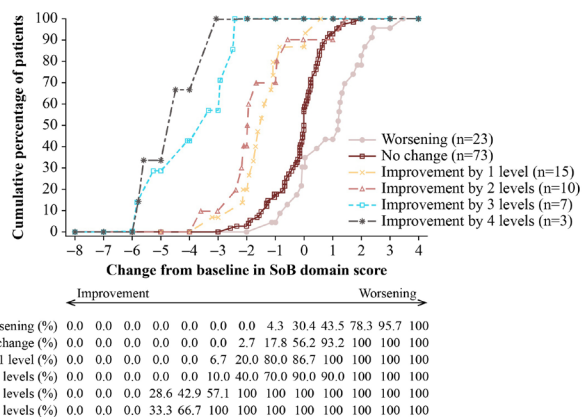


Figure 1 eCDF curves of changes in NTD-T-PRO T/W domain scores from baseline to weeks 13–24 by level of response on different anchors. *For the PGI-S, worsening, no change, improvement by 1 level, by 2 levels, by 3 levels and by ≥4 levels were defined as a change of ≥1, >−1 to <1, >−2 to ≤−1, >−3 to ≤−2, >−4 to ≤−3 and ≤−4 points, respectively. †For the FACIT-F FS, worsening, no change, improvement by 1 level and by 2 levels were defined as a change of ≤−4, >−4 to <4, ≥4 to <8 and ≥8 points, respectively. ‡For FACIT-F items HI12 and An2, worsening, no change, improvement by 1 level and by 2 levels were defined as a change of ≤−1, >−1 to <1, ≥1 to <2 and ≥2 to <3 points, respectively. §For SF-36v2 vitality, worsening, no change, improvement by 1 level and by 2 levels were defined as a change of ≤−6.7, >−6.7 to <6.7, ≥6.7 to <13.4 and ≥13.4 points, respectively. eCDF, empirical cumulative distribution function; FACIT-F, Functional Assessment of Chronic Illness Therapy–Fatigue; FS, Fatigue Subscale; NTD-T-PRO, non-transfusion-dependent beta-thalassaemia patient-reported outcome; PGI-S, Patient Global Impression of Severity; SF-36v2, Short Form Health Survey version 2; T/W, tiredness/weakness.

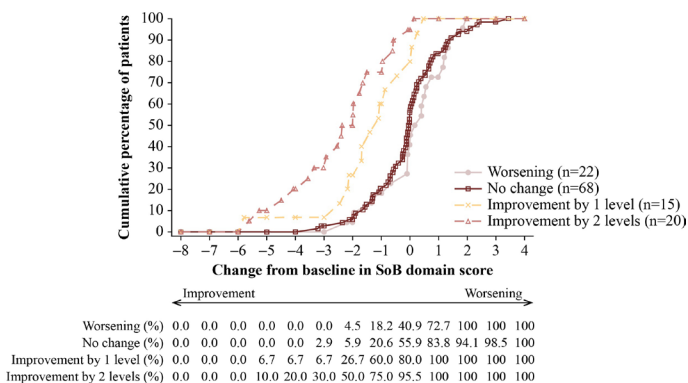
within-patient improvement threshold. The threshold was slightly less than the lower bound of the mean decreases in score for each of the improvements by one-level groups and was consistent with the optimal cut-off values from intersections of the PDF curves between the group with no change and the group with

one level of improvement (online supplemental figure S1) and the ROC analysis estimates. The threshold was also close to the 0.5 SD, but approximately twice the SEM (the amount of variability in the T/W score observed in the study participants that is caused by measurement error).

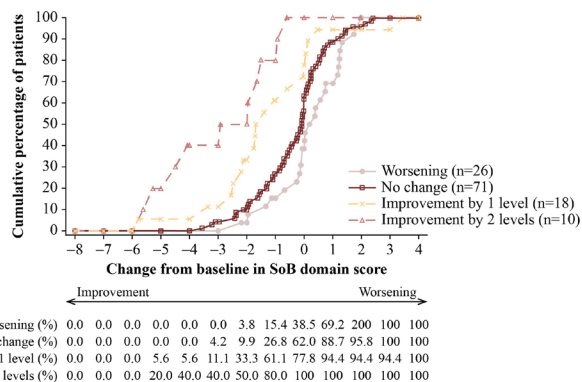
A PGI-S*



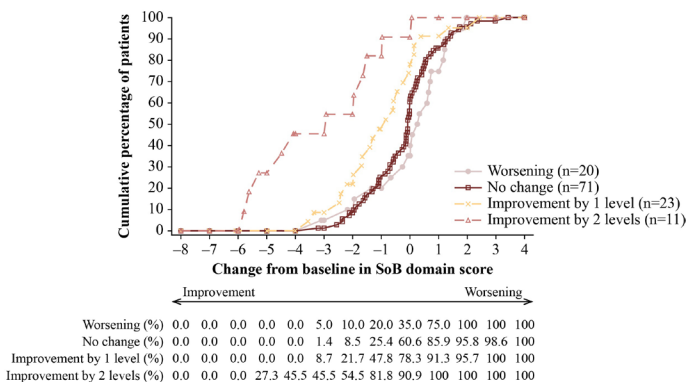
B FACIT-F FS†



C FACIT-F item HI12‡



D FACIT-F item An2‡



E SF-36v2 vitality§

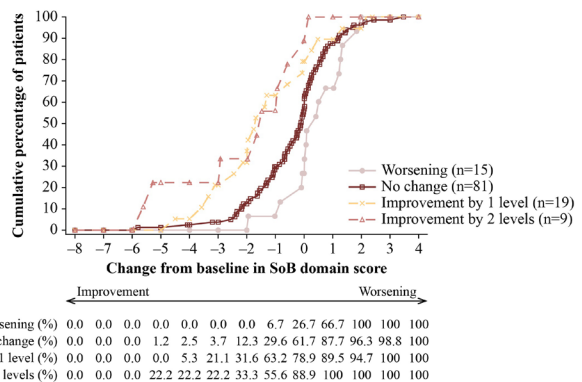


Figure 2 eCDF curves of changes in NTD-T-PRO SoB domain scores from baseline to weeks 13–24 by level of response on different anchors. *For the PGI-S, worsening, no change, improvement by 1 level, by 2 levels, by 3 levels and by ≥4 levels were defined as a change of ≥1, >−1 to <1, >−2 to ≤−1, >−3 to ≤−2, >−4 to ≤−3 and ≤−4 points, respectively. †For the FACIT-F FS, worsening, no change, improvement by 1 level and by 2 levels were defined as a change of ≤−4, >−4 to <4, ≥4 to <8 and ≥8 points, respectively. ‡For FACIT-F items HI12 and An2, worsening, no change, improvement by 1 level and by 2 levels were defined as a change of ≤−1, >−1 to <1, ≥1 to <2 and ≥2 to <3 points, respectively. §For SF-36v2 vitality, worsening, no change, improvement by 1 level and by 2 levels were defined as a change of ≤−6.7, >−6.7 to <6.7, ≥6.7 to <13.4 and ≥13.4 points, respectively. eCDF, empirical cumulative distribution function; FACIT-F, Functional Assessment of Chronic Illness Therapy–Fatigue; FS, Fatigue Subscale; NTD-T-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.

Among those patients with ≥1.0g/dL increase in mean haemoglobin values from baseline to weeks 13–24 (n=70/131, 53%), approximately 41% (29/70) also experienced a ≥1.0-point decrease in the T/W domain over the same time interval, compared with 30% (18/61) among those with <1.0g/dL increase in mean haemoglobin values.

Similarly, among those patients with ≥1.5g/dL increase in mean haemoglobin values from baseline to weeks 13–24 (n=44/131, 34%), 52% (23/44) also experienced a ≥1.0-point decrease in the T/W domain over the same time interval, compared with 28% (24/87) among those with <1.5g/dL increase in mean haemoglobin values.

Table 1 Changes in NTDT-PRO T/W domain scores by anchor response category from baseline to weeks 13–24

Anchor and statistic	Improvement				No change	Worsening by ≥ 1 level
	4 levels	3 levels	2 levels	1 level		
PGI-S	≤ -4	> -4 to ≤ -3	> -3 to ≤ -2	> -2 to ≤ -1	> -1 to < 1	≥ 1
N	3	7	10	15	73	23
Mean	-5.12	-4.08	-2.43	-2.14	-0.26	0.94
95% CI	-9.37 to -0.86	-5.50 to -2.66	-3.02 to -1.84	-2.58 to -1.71	-0.47 to -0.06	0.38 to 1.49
Median	-4.98	-3.56	-2.56	-1.98	-0.11	1.21
ES*	-2.33	-1.86	-1.11	-0.98	-0.12	0.43
FACIT-F FS	N/A	N/A	≥ 8	≥ 4 to < 8	> -4 to < 4	≤ -4
N	–	–	20	15	68	22
Mean	–	–	-2.95	-1.86	-0.25	0.37
95% CI	–	–	-3.75 to -2.14	-2.70 to -1.02	-0.54 to 0.04	-0.25 to 0.99
Median	–	–	-2.93	-1.76	-0.11	0.14
ES*	–	–	-1.34	-0.85	-0.11	0.17
FACIT-F item HI12	≥ 4	≥ 3 to < 4	≥ 2 to < 3	≥ 1 to < 2	> -1 to < 1	≤ -1
N	0	0	10	18	71	26
Mean	–	–	-3.51	-1.87	-0.53	0.41
95% CI	–	–	-5.03 to -1.98	-2.73 to -1.01	-0.83 to -0.23	-0.12 to 0.94
Median	–	–	-3.43	-2.06	-0.37	0.33
ES*	–	–	-1.60	-0.85	-0.24	0.19
FACIT-F item An2	≥ 4	≥ 3 to < 4	≥ 2 to < 3	≥ 1 to < 2	> -1 to < 1	≤ -1
N	0	0	11	23	71	20
Mean	–	–	-3.86	-1.31	-0.32	-0.03
95% CI	–	–	-5.20 to -2.51	-1.94 to -0.67	-0.61 to -0.03	-0.82 to 0.76
Median	–	–	-3.61	-1.47	-0.11	0.18
ES*	–	–	-1.76	-0.59	-0.15	-0.01
SF-36v2 vitality	N/A	N/A	≥ 13.4	≥ 6.7 to < 13.4	> -6.7 to < 6.7	≤ -6.7
N	–	–	9	19	81	15
Mean	–	–	-2.95	-1.68	-0.55	0.47
95% CI	–	–	-4.73 to -1.17	-2.55 to -0.82	-0.88 to -0.21	-0.07 to 1.00
Median	–	–	-2.46	-1.76	-0.29	0.72
ES*	–	–	-1.34	-0.77	-0.25	0.21

*ES was calculated as the mean change from baseline within the group divided by the overall SD of the NTDT-PRO T/W score at baseline. ES, effect size; FACIT-F, Functional Assessment of Chronic Illness Therapy–Fatigue; FS, Fatigue Subscale; N/A, not applicable; 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; T/W, tiredness/weakness.

SoB domain score

Similar to the T/W domain, the direction and magnitude of SoB mean and median changes from baseline over weeks 13–24 were consistent with the levels of improvement on each anchor (table 2). Mean changes from baseline (effect size) in the SoB domain scores for the group that improved by one level ranged from -0.93 (-0.41, FACIT-F item An2) to -1.40 (-0.61, PGI-S). Median changes in the anchors ranged from -0.83 (FACIT-F item An2) to -1.70 (SF-36v2 vitality). Distribution-based analyses gave estimates of 1.15 (0.5 SD) and 0.65 (SEM), with estimates from the ROC analyses indicating optimal

thresholds (maximising Youden's index) to be -0.62 (FACIT-F item An2), -0.87 (PGI-S and FACIT-F FS), -1.31 (SF-36v2 vitality) and -1.39 (FACIT-F item HI12). All ROC analyses exceeded the AUC threshold of 0.70, indicating good discriminant power.

Based on these findings, a ≥ 1.0 -point decrease in SoB score was also considered to represent a lower bound for the clinically meaningful within-patient improvement threshold based on the same justifications as stated for the T/W score.

Among patients with ≥ 1.0 g/dL increase in mean haemoglobin values from baseline to weeks 13–24

Table 2 Changes in NTDT-PRO SoB domain scores by anchor response category from baseline to weeks 13–24

Anchor and statistic	Improvement				No change	Worsening by ≥ 1 level
	4 levels	3 levels	2 levels	1 level		
PGI-S	≤ -4	> -4 to ≤ -3	> -3 to ≤ -2	> -2 to ≤ -1	> -1 to < 1	≥ 1
N	3	7	10	15	73	23
Mean	-4.40	-3.75	-1.59	-1.40	-0.14	0.94
95% CI	-7.54 to -1.26	-5.00 to -2.51	-2.55 to -0.63	-1.90 to -0.90	-0.34 to 0.07	0.43 to 1.45
Median	-4.49	-3.34	-1.97	-1.49	-0.01	1.19
ES*	-1.92	1.63	-0.69	-0.61	-0.06	0.41
FACIT-F FS	N/A	N/A	≥ 8	≥ 4 to < 8	> -4 to < 4	≤ -4
N	–	–	20	15	68	22
Mean	–	–	-2.37	-1.33	-0.06	0.20
95% CI	–	–	-3.14 to -1.61	-2.19 to 0.48	-0.36 to 0.23	-0.30 to 0.70
Median	–	–	-2.17	-1.09	-0.02	0.30
ES*	–	–	-1.03	-0.58	-0.03	0.09
FACIT-F item HI12	≥ 4	≥ 3 to < 4	≥ 2 to < 3	≥ 1 to < 2	> -1 to < 1	≤ -1
N	0	0	10	18	71	26
Mean	–	–	-2.90	-1.25	-0.32	0.27
95% CI	–	–	-4.22 to -1.58	-2.21 to -0.30	-0.62 to -0.03	-0.18 to 0.71
Median	–	–	-2.44	-1.53	-0.07	0.28
ES*	–	–	-1.26	-0.55	-0.14	0.12
FACIT-F item An2	≥ 4	≥ 3 to < 4	≥ 2 to < 3	≥ 1 to < 2	> -1 to < 1	≤ -1
N	0	0	11	23	71	20
Mean	–	–	-3.10	-0.93	-0.19	0.07
95% CI	–	–	-4.48 to -1.73	-1.55 to -0.32	-0.48 to 0.10	-0.56 to 0.69
Median	–	–	-2.92	-0.83	-0.07	0.29
ES*	–	–	-1.35	-0.41	-0.08	0.03
SF-36v2 vitality	N/A	N/A	≥ 13.4	≥ 6.7 to < 13.4	> -6.7 to < 6.7	≤ -6.7
N	–	–	9	19	81	15
Mean	–	–	-2.04	-1.34	-0.37	0.43
95% CI	–	–	-3.69 to -0.40	-2.19 to -0.49	-0.68 to -0.06	-0.13 to 1.00
Median	–	–	-1.49	-1.70	-0.12	0.41
ES*	–	–	-0.89	-0.58	-0.16	0.19

*ES was calculated as the mean change from baseline within the group divided by the overall SD of the NTDT-PRO SoB score at baseline. ES, effect size; FACIT-F, Functional Assessment of Chronic Illness Therapy–Fatigue; FS, Fatigue Subscale; N/A, not applicable; 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.

($n=70/131$, 53%), 39% (27/70) also experienced a ≥ 1.0 -point decrease in the SoB domain over the same time interval, compared with 26% (16/61) among those with < 1.0 g/dL increase in mean haemoglobin values. Similarly, among patients with ≥ 1.5 g/dL increase in mean haemoglobin values from baseline to weeks 13–24 ($n=44/131$, 34%), approximately 48% (21/44) also experienced a ≥ 1.0 -point decrease in the SoB domain over the same time interval, compared with 25% (22/87) among those with < 1.5 g/dL increase in mean haemoglobin values.

Clinically meaningful between-group difference threshold estimates

The clinically meaningful between-group difference threshold for the T/W domain was estimated to be in the range of 0.53–1.10 based on the SEM and 0.5 SD. Similarly, this threshold was estimated to be 0.65–1.15 for the SoB domain.

Symptomatic threshold for the NTDT-PRO T/W domain score

The optimal cut-off threshold to differentiate between patients who were symptomatic and those less/

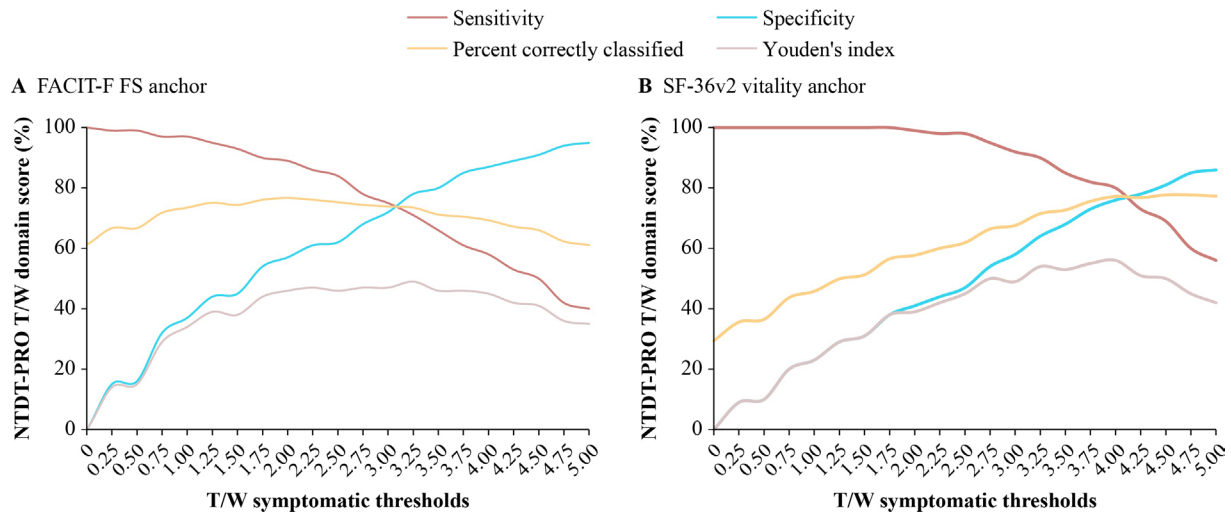


Figure 3 ROC analysis to identify a symptomatic threshold for the NTD-PRO T/W domain score. FACIT-F, Functional Assessment of Chronic Illness Therapy–Fatigue; FS, Fatigue Subscale; NTD-PRO, non-transfusion-dependent beta-thalassaemia patient-reported outcome; ROC, receiver operating characteristic; SF-36v2, Short Form Health Survey version 2; T/W, tiredness/weakness.

asymptomatic in fatigue, as identified by maximum Youden's index, was 3.04. The AUC (95% CI) for this analysis was 0.83 (0.80, 0.86), indicating very good discriminant power.³⁴ At this threshold, sensitivity and specificity were balanced (73% and 76%, respectively) and 74% of participants were classified correctly. As shown in figure 3A, sensitivity, specificity and per cent correctly classified curves intersect at approximately 3.

When using the SF-36v2 vitality domain with a threshold of <45 to differentiate patients who were symptomatic, a T/W domain score of 3.90 corresponded to the maximum Youden's index. The AUC (95% CI) indicated very good discriminant power (0.85 (0.82, 0.89)).³⁴ Sensitivity and specificity were 82% and 75%, respectively, and 77% of patients were specified correctly. The sensitivity, specificity and per cent correctly specified curves intersect at approximately 4 (figure 3B).

DISCUSSION

The current analysis aimed to determine a set of thresholds for the NTD-PRO to define meaningful within-patient change, interpret meaningfulness of between-group difference and identify patients with symptomatic T/W. Based on the triangulation of median and mean changes from baseline in the NTD-PRO scores in groups with improvement by one level on the selected anchors (tables 1 and 2), distribution-based estimates (ie, half of the SD of the baseline score and the SEM) and ROC curve analyses, ≥ 1.0 -point decrease was considered to represent a lower bound for the clinically meaningful within-patient improvement threshold on the NTD-PRO T/W or SoB score. This threshold aligns with estimates from the groups with one level of improvement on five anchors that measure similar concepts of interest and are correlated adequately with the T/W and SoB domains. It is also considerably higher than the calculated SEM

values, indicating that it is beyond the variability caused by measurement error, and consistent with the 0.5 SD estimate, commonly used as a good approximation of the clinically meaningful change thresholds for a given PRO measure.^{28 30} Additionally, the ≥ 1.0 -point reduction was demonstrated to appropriately reflect the proportions of patients with improvement in T/W and SoB among clinical responders and non-responders, as defined by haemoglobin improvements of 1.0 g/dL and 1.5 g/dL from baseline to weeks 13–24, in the BEYOND study. Using distribution-based methods, the thresholds for meaningful between-group differences were estimated to be in the range of 0.53–1.10 for the T/W domain and 0.65–1.15 for the SoB domain. These between-group difference thresholds reflect a small to medium effect size of treatment effect, consistent with the between-group minimally important differences of 0.3 SD and 0.5 SD reported for the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire Core 30 (QLQ-C30), as well as the treatment effect size assumption used to estimate the statistical power for the BEYOND study.^{16 36}

The cut-off threshold for distinguishing symptomatic from less/asymptomatic patients on the NTD-PRO T/W domain score was estimated to be between 3 and 4 points, based on ROC analysis using the FACIT-F FS as an anchor. While the estimate derived using the SF-36v2 vitality score as an anchor was approximately 1 point higher, a threshold of 3 points was ultimately chosen due to the higher reliability of the FACIT-F FS symptomatic anchor. Specifically, the threshold of 43 on the FACIT-F FS was identified using ROC curve analysis based on data from a large cohort study composed of anaemic patients with cancer (n=2369), non-anaemic patients with cancer (n=113) and the general US population (n=1010).³² The 43-point threshold best distinguished anaemic cancer

patients from the general population and classified study participants into the correct group 84% of the time, with a sensitivity of 0.92 and specificity of 0.69.³² On the other hand, the 45-point threshold from the SF-36v2 vitality score is proposed by the instrument developers as a general threshold to distinguish between the ‘average’ or ‘normal’ range for the US general population based on half of the SD of T-score but has not been formally validated.²⁰

The study findings have important practical implications, allowing clinical researchers to identify patients with meaningful improvements in NTDT symptoms when assessing treatment effects and evaluate the meaningfulness of differences in mean NTDT-PRO scores between treatments, in the context of a clinical trial.

Strengths, limitations and generalisability of this study

These are the first publicly reported thresholds to interpret the NTDT-PRO domain scores. One of the strengths of the present study is the use of multiple anchors, which measure similar concepts of interest from patients’ perspectives and are adequately correlated with the T/W and SoB domains, to determine the threshold for meaningful within-patient improvement in the NTDT-PRO. Furthermore, multiple methods, including mean and median NTDT-PRO scores within anchor categories, eCDF and PDF curves, ROC curve analysis and distribution-based analyses, were used to derive the estimates of the possible thresholds for triangulation. Finally, data used in these analyses were collected in a representative interventional study that spanned across several geographic regions and included patients with a wide range of symptom severity.

Limitations of this study include the use of non-VRS anchors (PGI-S, FACIT-F FS and SF-36v2 vitality), which is not consistent with FDA guidance,¹⁷ as a priori-determined groups for these anchors could be difficult to define. However, the estimates derived from these non-VRS anchors were consistent with those from the VRS anchors (FACIT-F items HI12 and An2), indicating that the cut-off values used to define a priori-determined groups on these anchors should be appropriate. As there is still no recommended guidance on how the threshold of meaningful between-group difference for a given PRO measure should be derived, the proposed ranges of threshold for the meaningful between-group difference were based on the distribution-based methods, which may vary if different analysis populations are considered and/or different effect sizes are assumed. Nevertheless, our estimates are consistent with the meaningful between-group difference thresholds for the EORTC QLQ-C30 domains estimated from the anchor-based approach, which were reported to mostly fall between small (0.3 SD) and medium (0.5 SD) effect size.³⁶ Finally, the cut-off threshold to discriminate between less/asymptomatic and symptomatic patients on the T/W domain score was based on the optimal threshold (43 points) of the FACIT-F FS to differentiate anaemic patients with cancer

from the general population.³² As such, the cut-off of ≥ 3 points as the symptomatic threshold for the NTDT-PRO T/W score proposed by the current analysis should be further validated using better anchors directly assessed from the same target population.

Regarding the generalisability of the study findings, while NTDT-PRO has been used effectively in a clinical trial, it has not been tested in routine clinical practice. Nevertheless, it holds potential for real-world application, enabling clinicians to identify patients requiring symptom relief and to assess treatment benefits. Further evaluation is necessary to determine the effectiveness of NTDT-PRO as a single-use assessment during clinical visits. Currently, the tool is validated for daily use with a 24-hour recall period, and its utilisation as a one-time assessment with a longer recall period may not be appropriate.

Conclusions

A ≥ 1 -point decrease in NTDT-PRO T/W and SoB domain scores represents a meaningful improvement from baseline to weeks 13–24 for patients with NTDT. Between-group difference thresholds were estimated to be in the range of 0.53–1.10 for the T/W domain and 0.65–1.15 for the SoB domain. A 3-point threshold can be used to distinguish between symptomatic and less/asymptomatic patients on the T/W domain score. These thresholds may be useful in future interventional or observational studies in NTDT to assess and interpret treatment effects over time, as well as help identify patients who need symptom relief.

Author affiliations

- ¹Department of Internal Medicine, American University of Beirut Medical Center, Beirut, Lebanon
- ²Center for Research on Rare Blood Disorders (CR-RBD), Burjeel Medical City, Abu Dhabi, UAE
- ³Division of Hematology/Oncology, Department of Pediatrics, Weill Cornell Medicine, New York city, New York, USA
- ⁴Siriraj Research Hospital, Mahidol University, Bangkok, Thailand
- ⁵First Department of Paediatrics, National and Kapodistrian University of Athens, Athens, Greece
- ⁶Bristol Myers Squibb, Princeton, New Jersey, USA
- ⁷Evidera, Waltham, Massachusetts, USA
- ⁸Evidera, Atlanta, Georgia, USA
- ⁹Adelphi Values, Boston, Massachusetts, USA
- ¹⁰Celgene International Sàrl, a Bristol–Myers Squibb Company, Boudry, Switzerland
- ¹¹Bristol Myers Squibb, Madison, New Jersey, USA
- ¹²Fondazione IRCCS Ca’ Granda Policlinico Hospital, University of Milan, Milan, Italy

Acknowledgements Athanasia Benekou and Stephen Gilliver, PhD of Evidera provided medical writing support and Eilish McBurnie, PhD of Excerpta Medica provided editorial support in accordance with Good Publication Practice guidelines (Good Publication Practice (GPP) Guidelines for Company-Sponsored Biomedical Research: 2022 Update | Annals of Internal Medicine (acpjournals.org)), funded by Bristol Myers Squibb.

Contributors JL-B, AY, SG, CGP and ALS contributed to protocol development. SG, CGP and ALS made substantial contributions to the design and concept of the study. ATT, VV, AK and MDC contributed to data acquisition. SG and CGP conducted the data and statistical analyses. ATT, KMM, VV, AK, JL-B, AY, SG, CGP, ALS, JKS, MBG, 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 accepted full



- Individual Changes in Health-Related Quality of Life. *J Clin Epidemiol* 1999;52:861–73.
- 28 Wyrwich KW. Minimal Important Difference Thresholds and the Standard Error of Measurement: Is There a Connection? *J Biopharm Stat* 2004;14:97–110.
- 29 Rejas J, Pardo A, Ruiz MÁ. Standard error of measurement as a valid alternative to minimally important difference for evaluating the magnitude of changes in patient-reported outcomes measures. *J Clin Epidemiol* 2008;61:350–6.
- 30 Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 2003;41:582–92.
- 31 Nunnally JC, Bernstein IH. *Psychometric Theory*. 3rd edn. New York, NY: McGraw-Hill, 1994.
- 32 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.
- 33 Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd edn. New York, NY: John Wiley & Sons, 2000.
- 34 Bekkar M, Djemaa H, Alitouche TA. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl* 2013;3:27–38.
- 35 Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
- 36 Musoro JZ, Coens C, Sprangers MAG, *et al*. Minimally important differences for interpreting EORTC QLQ-C30 change scores over time: A synthesis across 21 clinical trials involving nine different cancer types. *Eur J Cancer* 2023;188:171–82.