PEER REVIEW HISTORY

ARTICLE DETAILS

Title (Provisional)

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

Authors

Taher, Ali T.; Musallam, Khaled M.; Viprakasit, Vip; Kattamis, Antonis; Lord-Bessen, Jennifer; Yucel, Aylin; Guo, Shien; Pelligra, Christopher; Shields, Alan L.; Shetty , Jeevan K; Glassberg, Mrudula B; Bueno, Luciana Moro; Cappellini , MD

VERSION 1 - REVIEW

Reviewer	**1**
Name	**Garg, Akanksha**
Affiliation	**The Gujarat Cancer and Research Institute**
Date	**01-Mar-2024**
COI	**None**

The authors have proposed the use of the NTDT-PRO symptom assessment scale for patients who have responded to treatment in the BEYOND study. The objectives, statistical analysis and results have been discussed in detail.The strengths and limitations of the study have been described well.

Minor query

Q.1 Please comment on the use of this symptom assessment tool in the real world

Thanks

Reviewer	**2**
Name	**Foong, Wai Cheng**
Affiliation	**RCSI & UCD Malaysia Campus**
Date	**07-Jul-2024**

COI	**None**

An important study.

VERSION 1 - AUTHOR RESPONSE

Manuscript ID: bmjopen-2024-085234.R1

Please note that page and line numbers indicated in the responses below correspond to those visible in the clean version of the manuscript.

Reviewer #1

Dr. Akanksha Garg, The Gujarat Cancer and Research Institute

Comments to the Author:

The authors have proposed the use of the NTDT-PRO symptom assessment scale for patients who have responded to treatment in the BEYOND study. The objectives, statistical analysis and results have been discussed in detail. The strengths and limitations of the study have been described well.

1. Minor query: Please comment on the use of this symptom assessment tool in the real world. Thanks.

Response: Thank you for the positive feedback. We have expanded the "Strengths, limitations and generalisability of this study" subsection to include a comment on the real-world applicability of the NTDT-PRO (lines 405–411).

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 benefit. 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.

Reviewer #2

Dr. Wai Cheng Foong, RCSI & UCD Malaysia Campus

Comments to the Author:

An important study. My comments are in the attached file (listed below with author responses):

1. […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…] (pg 5)

Comment:

- suggest to remove

Response: Thank you for your suggestion. We have retained the word "double-blind" to ensure that the language remains aligned with that reported for the BEYOND trial.

2. […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)…] (pg 9)

Comment:

- [A] Would appreciate if the authors would consider to giving a brief description of the randomisation that took place (instead of the reader needing to search for the other paper).

Response: Thank you for your comment. We have revised the Methods section – Study design and participants – and added the following brief description of the randomisation procedure used in BEYOND (lines 130–133):

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).

- [B] Would appreciate the authors highlighting the following for each NTDT-PRO outcome assessed (which are rather subjective): i) who is blinded?, ii) were the participants aware of their assigned intervention?, iii) were the investigators and assessors aware of the participant' assigned intervention?, iv) any deviations from the intended intervention (effect of assignment to intervention and or effect of adhering to the intervention), v) was intention-to-treat analysis used?, vi) Could the results of the assessment for each outcome been influenced by knowledge of intervention? Would like to suggest to remove the word double blind on the abstract. Suggest to explain how blinding was made possible and how successful was the blinding implemented for each outcome at Results section.

Response: Thank you for your comments. We have added the following description of the study blinding process to the manuscript (lines 133–137). Additional details of the analyses have not been added to the manuscript, given the word count limitations, but are described elsewhere (Taher et al. 2022: Luspatercept for the treatment of anaemia in non-transfusion-dependent β-thalassaemia (BEYOND): a phase 2, randomised, double-blind, multicentre, placebo-controlled trial; Taher et al. 2023: Psychometric evaluation of the NTDT-PRO questionnaire for assessing symptoms in patients with non-transfusion-dependent beta-thalassaemia | BMJ Open).

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.

3. […Assessments…] (pg 9)

Comment:

It has been challenging to fully grasp the necessity and interrelation of the multiple tools used. While the individual scores are well-explained, I am particularly interested in understanding how these tools collectively contribute to the evaluation and how you determined the analyses indicate good discriminant power.

- Please explain: i) The rationale behind the selection of multiple measuring tools and how they complement each other in this study, ii) The method used to determine good discriminant power, particularly in relation to the ROC analyses and the significance of the AUC threshold exceeding 0.70.

Response: Thank you for your comments, which we have addressed as follows:

i) We added a sentence (line 171) to explain that multiple anchors were used to comply with US FDA guidance, which recommends that thresholds be derived from multiple anchors/methods (anchor-based and distribution-based) and then triangulated to determine the meaningful threshold or a range of thresholds.

We also clarified that the PRO measures and clinical outcomes we considered as anchors in the present study (other than NTDT-PRO) were those used alongside NTDT-PRO in the BEYOND study (lines 171–173).

For each PRO, the specific statements/questions that were considered for anchors, and the rationale for selecting them, are provided on lines 173–188 of the manuscript.

A detailed description of the various PROs used alongside NTDT-PRO and how they complement each other, in terms of the specific outcomes they measure, is provided in the Supplement (as now mentioned on lines 157–159 in the main body of the manuscript).

ii) As stated in the "Clinically meaningful within-patient threshold for improvement" section (lines 164–165 in the manuscript), in line with FDA guidance, we used anchor-based analyses as the primary approach, and these were supported by distribution-based analyses (i.e., half of the standard deviation of the baseline score and the standard error of measurement, which indicates the lower bound of the thresholds as they show the variability of the measure).

In the "Symptomatic threshold" section, we added details to clarify how the symptomatic threshold for the NTDT-PRO T/W score was estimated. Specifically, we explained that the FACIT-F FS (comprising 13 items specific to fatigue) and SF-36v2 vitality (comprising questions about patients' perception of their energy levels and tiredness) were selected because their concepts overlap with those that the NTDT-PRO T/W aims to capture (ie, tiredness and weakness) (lines 231–235).

Additionally, we expanded on the ROC analyses and AUC values considered to determine good discrimination power for NTDT-PRO T/W score. We explained that, for ROC analyses, AUC values of 0.5, 0.7 and 1.0 indicate no diagnostic ability (ie, similar to random guessing), good diagnostic ability and perfect diagnostic accuracy, respectively (lines 238–240). We therefore chose an AUC exceeding 0.7 as the threshold to evaluate how well the NTDT-PRO T/W score could discriminate between symptomatic and less/asymptomatic patients.

4. […Clinically meaningful within-patient improvement threshold estimates…] (pg 13)

Comment:

• For each domain scores, the authors did a commendable job in describing the results, but to someone without an extensive knowledge background, some of the explanations quite challenging to comprehend. It would be highly beneficial to present the key findings in a simpler manner that retains their clarity and significance. Suggest a more straightforward representation of the results for lay audience to enhance understanding and appreciation of this effort.

Response: Thank you for your feedback. Given the technical nature of the subject matter, the level of detail is a key strength of the manuscript, as we provide clear justification for the choices taken at each step, to ensure objectivity throughout. To enhance clarity for a non-expert audience, we have expanded the Discussion section to include the following description and added a paragraph explaining the implications of the study findings.

Lines 341–345:

Based on the triangulation of median and mean changes from baseline in the NTDT-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 NTDT-PRO T/W or SoB score.

Lines 374–377:

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

5. […Median changes in the anchors ranged from –0.83 (FACIT-F item An2) to –1.70 (SF-36v2 vitality)….] (pg 15)

Comment:

• Why is median changes used instead of mean changes? Please explain how heavily were the scores influenced by outliers and skewed data, which may lead to misleading representation.

Response: We have provided both means and medians in the manuscript (in the text and tables) and they are closely aligned (within 0.5 of each other in almost all instances), indicating minimal data skewness. The median is generally preferred when deriving meaningful within-patient change thresholds, as it reflects a change score that can actually occur in an individual; it represents the 50th percentile and therefore the actual score of a patient if the number of observations is odd, or the mean of the two middle scores if the number of observations is even. The same is not true for the mean. This is particularly important for instruments that use a transformed scale, though it is not a critical issue for NTDT-PRO, which is not a transformed scale.

6. […Handling of missing PRO data…] (pg 32)

Comment:

• [A] What are the reasons that the outcome data were missing (for each outcome)? Are the reasons explained somewhere in the text? These information are important because most of the outcomes were on participants' perceptions which are rather subjective.

Response: Among the ITT population, completion rate of the NTDT-PRO was high, ranging from 100% at baseline to 87.3% at week 24 (supplemental figure 1 in Taher et al. 2023). At least 124 patients (83.2%) were included in the anchor-based analyses, having a non-missing NTDT-PRO domain score and a non-missing anchor score of interest at both baseline and weeks 13–24. We do not have specific reasons for missingness for each outcome; however, given the high proportion of patients included in the analyses, missing data are unlikely to have a substantial impact on the results, and the analyses can be considered representative of the intention-to-treat (ITT) population. Furthermore, in analyses evaluating meaningful thresholds on a PRO with a focus on relationships between various PRO measures (as opposed to an analysis of treatment effect), missing data are less of a concern. Additionally, FDA guidance for determining within-patient threshold states that only “empirical” data should be considered. Estimates from modelling approaches (including missing imputation) are not preferred (ie, why we use the empirical cumulative distribution function [eCDF]).

• [B] Is there evidence that the results for each outcome was not biased by missing outcome data? How were the missing data handled? Any sensitivity analyses done to assess any plausible values of the missing outcome data? Please elaborate.

Response: The goal of the manuscript was not to estimate the impact of treatment on the PROs, but to derive meaningful thresholds for the NTDT-PRO. Therefore, we did not attempt to impute any missing data. As mentioned above, this is based on FDA guidance stating a preference for empirical data, rather than the use of models or imputation. This rationale has been clarified in the Supplement (Supplemental material; lines 37–39). Given the high compliance rate, the analyses can be considered representative of the ITT population.

An analysis was conducted to confirm the weekly scoring rule for NTDT-PRO items (ie, a weekly score was considered missing if scores were available for <4 days) by varying the number of missing days allowed to calculate weekly score (1, 2, 3, 4, 5 or 6) scenarios, and by calculating the mean/SD scores across patients for each missing day scenario. The mean and SDs were similar for each missing day scenario, indicating that the scoring rule was appropriate and the scores did not vary substantially with missing data. Details of these analyses have been previously published (Taher et al. 2023).

VERSION 2 - REVIEW

Reviewer	**2**
Name	**Foong, Wai Cheng**
Affiliation	**RCSI & UCD Malaysia Campus**
Date	**28-Sep-2024**
COI	

Thank you for the clarifications, addressing my comments and making the necessary revisions. I have no further feedback. Nothing to add.