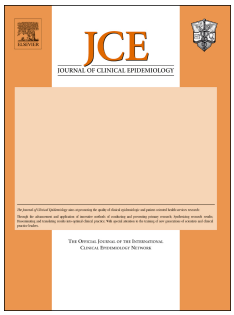

Thresholds for clinical importance were defined for the EORTC CAT Core – an adaptive measure of core quality of life domains in oncology clinical practice and research

Johannes M. Giesinger, Fanny L.C. Loth, Neil K. Aaronson, Juan I. Arraras, Giovanni Caocci, Fabio Efficace, Mogens Groenvold, Marieke van Leeuwen, Morten Aa Petersen, John Ramage, Krzysztof A. Tomaszewski, Teresa Young, Bernhard Holzner, on behalf of the EORTC Quality of Life Group

PII: S0895-4356(19)30621-3

DOI: https://doi.org/10.1016/j.jclinepi.2019.09.028

Reference: JCE 9990

To appear in: *Journal of Clinical Epidemiology*

Received Date: 18 July 2019

Revised Date: 17 September 2019

Accepted Date: 30 September 2019

Please cite this article as: Giesinger JM, Loth FLC, Aaronson NK, Arraras JI, Caocci G, Efficace F, Groenvold M, van Leeuwen M, Petersen MA, Ramage J, Tomaszewski KA, Young T, Holzner B, on behalf of the EORTC Quality of Life Group, Thresholds for clinical importance were defined for the EORTC CAT Core – an adaptive measure of core quality of life domains in oncology clinical practice and research, *Journal of Clinical Epidemiology* (2019), doi: https://doi.org/10.1016/j.jclinepi.2019.09.028.

Thresholds for clinical importance were defined for the EORTC CAT Core – an adaptive measure of core quality of life domains in oncology clinical practice and research

Johannes M. Giesinger[a], Fanny L. C. Loth[a], Neil K. Aaronson[b], Juan I. Arraras[c], Giovanni Caocci[d], Fabio Efficace[e], Mogens Groenvold[f], Marieke van Leeuwen[b], Morten Aa Petersen[f], John Ramage[g], Krzysztof A. Tomaszewski[h], Teresa Young[i], Bernhard Holzner[a], on behalf of the EORTC Quality of Life Group

[a]University Hospital of Psychiatry II, Medical University of Innsbruck, Innsbruck, Austria

[b]Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

[c]Oncology Departments, Complejo Hospitalario of Navarre, Pamplona, Spain

[d]Department of Medical Sciences and Public Health, University of Cagliari, Cagliari, Italy

[e]Health Outcomes Research Unit, Italian Group for Adult Hematologic Diseases (GIMEMA) Data Center, Rome, Italy

[f]The Research Unit, Department of Palliative Medicine, Bispebjerg Hospital, and Department of Public Health, University of Copenhagen, Copenhagen, Denmark

[g]Department of Gastroenterology and Hepatology, Hampshire Hospitals NHS Foundation Trust, Aldermaston Road, Basingstoke, RG24 9NA, UK

[h]Faculty of Medicine and Health Sciences, Andrzej Frycz Modrzewski Krakow University, Krakow, Poland and Scanmed St. Raphael Hospital, Krakow, Poland

[i]Lynda Jackson Macmillan Centre, East & North Hertfordshire NHS Trust incorporating Mount Vernon Cancer Centre, Northwood, UK

Corresponding author:
Johannes M. Giesinger, PhD
University Hospital of Psychiatry II
Medical University of Innsbruck
Anichstrasse 35, 6020 Innsbruck, Austria
E-mail: johannes.giesinger@i-med.ac.at
Phone: +43 512 504 23691

Funding

The study was funded by a grant from the EORTC Quality of Life Group (grant number 008 2014). The Austrian Science Fund (FWF #P26930) funded the work of Johannes M. Giesinger.

Conflict of interests

Bernhard Holzner is an owner of the intellectual property rights of the software CHES. None of the other authors has a conflict of interest to declare.

Abstract

Objective

To establish thresholds for clinical importance (TCIs) for the EORTC Computer Adaptive Testing (CAT) Core measure, the new adaptive version of the EORTC QLQ-C30.

Study Design and Setting

For our diagnostic study, we recruited cancer patients with mixed diagnoses and treatments from six European countries. Patients completed the EORTC CAT Core and a questionnaire with anchor items assessing criteria for clinical importance (limitations in everyday life, need for help/care, and worries by the patient/family/partner) for each EORTC CAT Core domain. We used a binary variable summarizing the anchor items for determining TCIs and for calculating the Area under the Curve (AUC) in Receiving Operator Characteristic analysis as a measure of diagnostic accuracy.

Results

Using data from 498 cancer patients (mean age 60.4y, 55.2% women), we established TCIs for the 14 domains of the EORTC CAT Core. Median AUC across domains was 0.93 (range 0.84-0.94). Median sensitivity and specificity of the TCIs was 0.91 (range 0.80-0.96) and 0.77 (range 0.66-0.84), respectively. TCIs and AUCs were largely consistent across patient groups.

Conclusion

We have generated TCIs for the 14 functional health and symptom domains of the EORTC CAT Core. The EORTC CAT Core showed high diagnostic accuracy in identifying clinically important symptoms and functional impairments.

Running title: Clinical thresholds for the EORTC CAT Core
Keywords: quality of life, clinical oncology, patient-reported outcome measures, EORTC CAT Core, clinical significance, thresholds, cut-offs,
Word count: 3525

Introduction

In recent years, patient-reported outcomes (PROs) have become a cornerstone of clinical research and are increasingly being integrated into daily practice. This has been fostered by the availability of reliable and well-validated PRO measures that are derived from sophisticated development procedures [1-3].

PRO measures based on Item Response Theory (IRT), a probabilistic measurement theory to determine psychometric characteristics of PRO measures, have only recently been introduced in the medical field, although they have a long tradition of use in educational testing and psychological assessments [4, 5]. Measures developed according to IRT models are more versatile than traditional questionnaires that rely on classical test theory [6]. An important advantage of IRT-based PRO measures is that they allow one to administer different questions on a given topic to different patients, while still obtaining comparable scores across patients on the same metric [7].

Over the last few years, the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group (QLG) has developed item banks based on IRT models to enhance measurement for each of the functional health and symptom domains of the EORTC QLQ-C30 questionnaire. These so-called EORTC Computer Adaptive Testing (CAT) Core measures [8] consist of item banks with validated questions with well-defined measurement characteristics. The item banks allow for two modes of administration, CAT and static questionnaire short-forms.

CAT assessments rely on an algorithm [2] to tailor the questions to the individual patient, based on his/her earlier responses. Static short-forms are predefined sets of items from an item bank typically selected to maximize measurement precision for the score distribution of a specific patient population. While CAT requires electronic questionnaire administration (e.g. on a tablet PC or mobile phone), static short-forms can also be administered on paper. The EORTC CAT Core item banks have been shown to have higher measurement precision than the QLQ-C30, while providing scores that are compatible with the original QLQ-C30 [8].

Scores from the EORTC CAT Core are presented on a T-score metric, which is a standardized metric with a mean of 50 and a standard deviation of 10 points. Standardization is reached through use of a reference population [9]. This scoring method is different from the scoring of the QLQ-C30, which presents scores on a 0-100 metric derived from summing responses to individual questions and the use of linear transformation [10]. Since T-scores describe the difference from the reference population mean in standard deviation units, they are more informative than simple sum scores. While this facilitates

score interpretation, what constitutes a clinically important symptom or functional health impairment on such a metric is not well-defined.

Thresholds for clinical importance (TCIs) are needed to calculate prevalence rates for clinically important symptoms and functional impairments from metric EORTC CAT Core scores, and to improve applicability of this instrument for symptom screening and monitoring in daily clinical practice. Routine collection of PRO data in clinical practice has been demonstrated to improve symptom management and even to improve survival rates [11-14].

Such thresholds improve the interpretability of scores from individual patients at a single time point, which is conceptually different from minimal important [15, 16] used for evaluating score change over time or differences between patient groups.

In a previous analysis [17], we established TCIs for the scales of the EORTC QLQ-C30 and found excellent diagnostic accuracy for all scales as well as invariance of thresholds across various patient groups. The EORTC CAT Core covers the same functional health and symptom domains as the EORTC QLQ-C30, but provides more flexibility regarding assessment length (number of questions) and improved measurement precision [8, 18]. Combining these advantages with TCIs that facilitate score interpretation may make the EORTC CAT Core particularly useful for patient monitoring in daily practice.

The objectives of the current study were to establish TCIs for the EORTC CAT Core and to determine the sensitivity and specificity of these thresholds when used for identification of clinically important symptoms and functional health impairments.

Methods

Sample

For this cross-sectional study, we recruited cancer patients (any diagnosis, type of treatment, or treatment status) in six European countries (Austria, Italy, the Netherlands, Poland, Spain, and the United Kingdom). For inclusion in the study, patients had to be aged 18 or older, speak the primary language of their country of residence, and provide written informed consent. Patients were excluded from the study if they had serious cognitive impairment that would prohibit them from completing questionnaires. Patients completed the study questionnaires (a short-form based on the EORTC CAT Core and anchor items on clinical importance) either on paper or electronically via the software program CHES [19]. This analysis relies on the same data set that has been used previously for establishing TCIs for the EORTC QLQ-C30 [17].

Ethical approval was obtained from the local ethics committees, if required (Medical University of Innsbruck: AN-2014-0012; East & North Hertfordshire NHS Trust: IRAS code 145602; Netherlands Cancer Institute: METC-AVL P17TRE).

EORTC CAT Core measures

The EORTC CAT Core item banks comprise a total of 260 items (including all items from the QLQ-C30) organized into 14 item banks, each comprising between 7 and 34 items to cover one of the the five functional health and nine symptom domains of the EORTC QLQ-C30 questionnaire. The functional health domains comprise Physical Functioning (PF), Role Functioning (RF), Social Functioning (SF), Emotional Functioning (EF) and Cognitive Functioning (CF). The symptom domains cover Fatigue (FA), Pain (PA), Nausea/Vomiting (NV), Appetite Loss (AP), Dyspnoea (DY), Sleep Disturbances (SL), Diarrhoea (DI), Constipation (CO), and Financial Impact of Disease (FI). The EORTC CAT Core presents results as T-scores that are based on a representative general population sample of 11,343 individuals from 11 European countries [9]. Higher scores on the functioning scales indicate higher levels of functioning, while higher scores on the symptom scales represent more symptom burden.

In our study, we administered static short-forms that were created from the EORTC CAT Core item banks. Short-forms were used instead of adaptive assessments so that data could also be collected in a paper-pencil format. For each domain, we included the items from the QLQ-C30 and additional items from the item banks that increased measurement precision in the range where we expected, *a priori*, the TCIs to be located (i.e. between the 75^{th} to 90^{th} percentile of general population scores [20] for symptom domains, and between the 10^{th} and 25^{th} percentile for functional health domains) . In total, the short-forms comprised seven items for PF and EF, five items for FA and four items for all other domains.

Anchor items for establishing thresholds

In this study, we used the same anchor items as have been used previously to establish TCIs for the EORTC QLQ-C30. Details on the rationale for defining anchor items have been published elsewhere [17]. Briefly, following a mixed methods study in 150 cancer patients and health professionals [21] and a consensus meeting within the EORTC QLG, we defined clinical importance of a symptom or functional health impairment in terms of three criteria: limitations in everyday life, worries by the patient or his/her family/partner, and the need for help or care.

Specifically, the anchor questions used for establishing TCIs were the following:

- Limitations: 'Has your SYMPTOM/PROBLEM limited your daily life?'
- Worries: 'Has your SYMPTOM/PROBLEM caused you or your family/partner to worry?
- Need for help: 'Have you needed any help or care because of your SYMPTOM/PROBLEM?'

For RF, SF and PA, we did not ask about limitations, since interference with daily life is already included in the EORTC CAT Core for these three domains. In line with previous studies [17, 22], we used the standard QLQ-C30 response format for the anchor items (a 4-point Likert scale with responses choices "not at all", "a little", "quite a bit", and "very much"). A patient was categorized as a case (i.e. as having a clinically important problem/symptom) if (s)he selected 'quite a bit' or 'very much' on any of the anchor items. If neither of these two categories was selected, patients were categorized as non-cases.

Statistical analysis

Descriptive statistics for the EORTC CAT Core are given as means and standard deviations, separately for cases and non-cases, as defined above. Differences between the two groups are reported in terms of absolute differences on the T-score metric and as effect sizes (Cohen's d).

To investigate diagnostic accuracy and to establish TCIs we used receiver operating characteristic (ROC) analysis. In this analysis the binary variable (case/non-case) derived from the anchor items was used as the criterion and the EORTC CAT Core score as the predictor. In ROC analysis, the area under the curve (AUC) reflects how well a predictor variable discriminates between cases and non-cases. An AUC above 0.80 indicates excellent discrimination [23].

TCIs were determined based on the following stepwise decision rule that gave greater weight to sensitivity than to specificity: If possible, we selected a TCI providing maximum sensitivity with a specificity >0.80 (requiring the sensitivity to be >0.90). If such a TCI was not found, we selected a TCI with maximum sensitivity and a specificity >0.70 (requiring sensitivity to be >0.80). Finally, if no previous step allowed definition of a TCI, we selected a TCI with a sensitivity >0.80 and the highest achievable specificity.

We emphasized sensitivity over specificity since the main use of the TCIs will be for screening in daily practice, where under-identification of symptoms may be more problematic than "false alarms". As a sensitivity analysis of this decision we also calculated TCIs obtained by giving equal weight of sensitivity and specificity by calculating the Youden J

index (i.e. the sum of sensitivity and specificity minus 1 [24] and contrasting this index for our TCIs with the maximum obtainable Youden J value.

For each of the 14 domains of the EORTC CAT Core we investigated the robustness of diagnostic accuracy across various patient groups. For this purpose we calculated the AUC for 14 different patient groups, defined by: age (below/above 60y), sex, treatment intention (curative/palliative), treatment status (on/off), comorbidity (no/yes), and European region (Western Europe [Austria and the Netherlands], Southern Europe [Italy and Spain], Eastern Europe [Poland], and the UK). For each patient group we investigated if the AUC exceeded the threshold for excellent discrimination of 0.80.

For each domain we used a multivariate binary logistic regression model to evaluate the invariance of TCIs across these patient groups. The model included the above grouping variables and the EORTC CAT Core score as independent variables and the binary criterion variable (case/non-case) as dependent variable. In such a model the grouping variables indicate between-group differences regarding the probability of being a case for a specific EORTC CAT Core score, i.e. a difference in TCIs between groups. For statistically significant grouping variables ($p<0.01$), we investigated group-specific TCIs using the above decision rule within each patient group.

The sample size for this study was determined on the basis of an *a priori* power analysis for the ROC analysis. This analysis showed that a sample of 500 patients (assuming 33% cases) provides a power of 0.80 to demonstrate (with a two-sided alpha of 0.05) that the AUC is above 0.80 if the observed AUC is 0.865. The observed AUC was estimated based on results from a previous pilot study [22]. The power analysis was conducted with PASS 11.0 [25].

Results

Patient characteristics

Between November 2016 and November 2018 we recruited 502 patients, of whom 498 (mean age 60.4, SD 12.7; 55.2% women) provided complete questionnaires that could be used for the analysis. At the time of assessment, most patients (76.7%) were on-treatment (60.6% with curative intention). Further details are reported in Table 1 and elsewhere [17].

The percentage of cases on the EORTC CAT CORE domains, based on the criteria for clinical importance described above, ranged from 8.3% for DI to 54.5% for PF, with a median prevalence across domains of 18.1% (see Table 2).

Thresholds for clinical importance

For the functional health domains, we observed the largest difference between cases and non-cases for RF (44.5 vs 30.2 points, effect size (ES)=-1.76) and the smallest difference for SF (45.9 vs 35.6 points, ES=-1.34). For symptom scales differences ranged from an ES of 1.65 (SL: 50.1 vs 63.0 points) to an ES of 2.79 (NV: 51.5 vs 78.9 points). The median ES was -1.48 for the functioning scales, and 2.13 for the symptom scales. Further details are reported in table 2.

Diagnostic accuracy in terms of AUC was above 0.90 for 9 of the 14 scales. The largest AUCs were observed for FA, AP, CO, DI, FI (all AUC=0.94) and the lowest for PF and SF (both 0.84).

TCIs for functioning scales ranged from 37 points for RF to 46 points for PF and EF. For symptom scales the lowest TCI was observed for SL (55 points) and the highest for AP (63 points). Sensitivity of the TCIs ranged from 0.80 (SF) to 0.96 (CO), with a median value across domains of 0.91. Specificity was lowest for PF (0.66) and highest for Fatigue (0.84), with a median of 0.77. For further details please see table 3 and figure 1.

As mentioned above our study relied on static short-forms from the EORTC CAT Core item banks that were based on a priori assumptions about the TCIs. Higher measurement precision (i.e. item information) at the TCI allows for more accurate classification of patients with scores close to the TCI, resulting in an increase of the sensitivity and specificity of the scale. For an illustration of how measurement precision at the TCI differs, please see figure 2 showing as an example the five-item QLQ-C30 Physical Functioning scale and a five-item static short-from designed to maximize measurement precision at the TCI.

Sensitivity analysis

Investigating group-specific AUCs in 14 patient groups for the 14 functional health and symptom domains we found that only 5 of 196 AUCs were below 0.80 (the threshold for excellent discrimination): PF in the UK (AUC=0.77), CF in Southern Europe (0.77), EF in Western Europe (0.78), PF in Western Europe (0.79), and CF in patients on-treatment (0.79). The 95% confidence intervals of all of these five AUCs included 0.80.

Our sensitivity analysis of the robustness of TCIs across patient groups using a logistic regression model indicated statistically significant ($p<0.01$) differences in TCIs for 6 of the 196 combinations of patient groups and domains. When applying the above decision rule for determining TCIs to these individual patient groups and domains, we found that the overall TCI differed by more than one point from the group-specific TCI for the following scales: PF in patients <60 years, TCI=48 (sens. 0.80, spec. 0.64); PF in patients >=60 years,

TCI=44 (sens 0.87, spec 0.73); SF in Southern Europe, TCI=43 (sens. 0.86, spec. 0.62); FA in Eastern Europe, TCI=55 (sens. 0.95, spec. 0.90); DY in the UK, TCI=62 (sens. 0.96, spec. 0.84); and DY in Eastern Europe, TCI=58 (sens. 0.91, spec. 0.73). For all these domains the difference between the group-specific and the overall TCI was 2 points.

Comparing the Youden J index for the TCIs derived from our decision rule against the maximum obtainable value, we found a difference exceeding 0.05 for three domains: For PF a threshold of 44 provides a Youden J index that is larger by 0.077, for EF a threshold of 42 increases Youden J by 0.061, and for CO a threshold of 63 has a Youden J higher by 0.074.

Discussion

We have established TCIs for all domains of the EORTC CAT Core, the adaptive PRO instrument recently developed by the EORTC Quality of Life Group. We found excellent diagnostic accuracy for the EORTC CAT Core measures in identifying clinically important functional health impairments and symptoms, which facilitated defining TCIs with high sensitivity and, in general, high specificity. TCIs were in the range of 4 to 13 points (i.e. 0.4 to 1.3 standard deviation units) from the normative general population mean of 50 points. The sensitivity analysis indicated that, with very few exceptions, the diagnostic accuracy of the EORTC CAT Core measure was excellent for the patient groups analyzed. Evaluating the performance of the TCIs in specific patient groups, we found minor differences in the optimal TCI for a small number of combinations of domains and patient groups, most notably a smaller impairment in PF being of clinical importance in patients below 60 years, compared to those above 60 years. In general, the EORTC CAT Core outperformed the EORTC QLQ-C30 in terms of diagnostic accuracy [17] resulting in better sensitivity and specificity of the TCIs. This makes the use of the EORTC CAT Core attractive in daily clinical practice, where high measurement precision is desirable at the individual patient level. Diagnostic accuracy may be improved further by relying on computer-adaptive assessments that maximize measurement precision of the EORTC CAT Core for scores close to the TCIs. For centers relying on paper-pencil data collection or using software not capable of administering CAT measures, static short-forms can be created that maximize measurement precision around the TCI to allow for accurate identification of clinically important problems (please see Figure 2).

With an intention similar to our study, a series of studies have established thresholds for severity categories for several PROMIS measures [26-28]. However, these studies used a quite different methodological approach, relying on case vignettes describing a range of

possible severity levels for each domain. The case vignettes were created based on item content and responses and described symptom levels that each differed by 5 points on a T-score metric (i.e. by 0.5 standard deviations). These case vignettes were then ranked and categorized as describing normal, mild, moderate or severe symptom levels by clinicians and patients [27, 28] or by clinicians only [26]. Comparing classifications by patients and clinicians, thresholds were fairly consistent, with patients sometimes rating symptom descriptions as more severe [27] and sometimes as less severe [28].

While the PROMIS measures and the EORTC CAT Core both present scores on a T-score metric, comparability of the thresholds is limited due to differences between the measures in terms of content and measurement characteristics, the normative population underlying the scoring (US vs Europe), and the methodology used to establish the thresholds. Also, a short-coming of the case vignette method is that the criteria for setting the thresholds are not as explicit as in our study and the sensitivity and specificity for the established thresholds are not available, thus not allowing a comparison with our TCIs in this regard. However, consistent with our study, there was also substantial variation in thresholds across domains for the PROMIS measures and for certain domains (e.g. pain, fatigue [26]) thresholds were closer to the normative mean of 50 than one might expect. In fact, for some of the PROMIS measures the general population mean overlapped with the categories for mild symptom levels [27, 28]. This could indicate a response shift phenomenon [29] resulting in an underestimation of the true difference between cancer patients and the general population, but may also reflect the high percentage of individuals suffering from (chronic) diseases in the general population. In the normative sample for the EORTC CAT Core [9], for example, 61.0% of the participants from the general population reported at least one health condition, with chronic pain (23%), arthritis (13%) and diabetes (10%) being most common.

Development of thresholds for PRO measures has been recommended in the literature, because the interpretation of scores on abstract metrics has been identified as one of the major barriers to the use of PRO measures in daily clinical practice [30]. The TCIs for the EORTC CAT Core can be integrated into software used for routine PRO monitoring to improve graphical presentation of PRO results (e.g. use of color-coding or reference lines [31, 32]. In addition, TCIs make PRO scores more actionable and support the linking of PRO results to clinical decision making [33, 34]. Our results provide a key component for the successful implementation of the EORTC CAT Core into daily clinical practice, where its flexibility and measurement precision at the individual patient-level may be particularly important.

A limitation of our study is that we have not used CAT assessments, but short-forms that were created to maximize measurement precision in the range where we expected the TCIs. With CAT assessments or short-forms targeting the now known TCIs more precisely, a higher diagnostic accuracy may be obtained, implying that the AUCs reported in our study may actually underestimate the diagnostic accuracy obtainable with the EORTC CAT Core measure.

A strength of our methodological approach is that we were able to relate the thresholds to explicit criteria that have been developed carefully, relying on interviews with patients and health care professionals, as well as on input from PRO experts in the EORTC QLG [17, 21]. The clear definition allows for a better understanding of the actual meaning of the thresholds. Furthermore, our empirical approach allowed us to estimate the sensitivity and specificity of the TCIs and to conduct a detailed analysis of invariance across patient groups.

In conclusion, we have established TCIs for the EORTC CAT Core measure that will facilitate the use of this measure for PRO monitoring in clinical practice. In clinical research, the TCIs may be used for converting metric T-scores to symptom prevalence rates that may be easier to interpret.

References

1. Sprangers, M.A., A. Cull, K. Bjordal, M. Groenvold, and N.K. Aaronson, *The European Organization for Research and Treatment of Cancer. Approach to quality of life assessment: guidelines for developing questionnaire modules. EORTC Study Group on Quality of Life.* Qual Life Res, 1993. **2**(4): p. 287-95.
2. Petersen, M.A., M. Groenvold, N.K. Aaronson, W.C. Chie, T. Conroy, A. Costantini, P. Fayers, J. Helbostad, B. Holzner, S. Kaasa, S. Singer, G. Velikova, and T. Young, *Development of computerised adaptive testing (CAT) for the EORTC QLQ-C30 dimensions - general approach and initial results for physical functioning.* Eur J Cancer, 2010. **46**(8): p. 1352-8.
3. Reeve, B.B., R.D. Hays, J.B. Bjorner, K.F. Cook, P.K. Crane, J.A. Teresi, D. Thissen, D.A. Revicki, D.J. Weiss, R.K. Hambleton, H. Liu, R. Gershon, S.P. Reise, J.S. Lai, and D. Cella, *Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS).* Med Care, 2007. **45**(5 Suppl 1): p. S22-31.
4. Wright, B.D. and M.H. Stone, *Best test design.* 1979.
5. Rasch, G., *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* 1960.
6. DeVellis, R.F., *Classical test theory.* Medical care, 2006: p. S50-S59.
7. Van Der Linden, W. and R. Hambleton, *Handbook of Item Response Theory*. Vol. 3. 2016: CRC Press Inc.
8. Petersen, M.A., N.K. Aaronson, J.I. Arraras, W.C. Chie, T. Conroy, A. Costantini, L. Dirven, P. Fayers, E.M. Gamper, J.M. Giesinger, E.J.J. Habets, E. Hammerlid, J. Helbostad, M.J. Hjermstad, B. Holzner, C. Johnson, G. Kemmler, M.T. King, S. Kaasa, J.H. Loge, J.C. Reijneveld, S. Singer, M.J.B. Taphoorn, L.H. Thamsborg, K.A. Tomaszewski, G. Velikova, I.M. Verdonck-de Leeuw, T. Young, M. Groenvold, R. European Organisation for, and G. Treatment of Cancer Quality of Life, *The EORTC CAT Core-The computer adaptive version of the EORTC QLQ-C30 questionnaire.* Eur J Cancer, 2018. **100**: p. 8-16.
9. Liegl, G., M.A. Petersen, M. Groenvold, N.K. Aaronson, A. Costantini, P.M. Fayers, B. Holzner, C.D. Johnson, G. Kemmler, K.A. Tomaszewski, A. Waldmann, T.E. Young, M. Rose, S. Nolte, and E.Q.o.L. Group, *Establishing the European Norm for the health-related quality of life domains of the computer-adaptive test EORTC CAT Core.* Eur J Cancer, 2019. **107**: p. 133-141.
10. Fayers, P.M., N.K. Aaronson, K. Bjordal, D. Curran, and M. Grønvold, *EORTC QLQ-C30 scoring manual*1999: Eortc.
11. Basch, E., A.M. Deal, A.C. Dueck, H.I. Scher, M.G. Kris, C. Hudis, and D. Schrag, *Overall Survival Results of a Trial Assessing Patient-Reported Outcomes for Symptom Monitoring During Routine Cancer Treatment.* JAMA, 2017. **318**(2): p. 197-198.
12. Denis, F., E. Basch, A.L. Septans, J. Bennouna, T. Urban, A.C. Dueck, and C. Letellier, *Two-Year Survival Comparing Web-Based Symptom Monitoring vs Routine Surveillance Following Treatment for Lung Cancer.* JAMA, 2019. **321**(3): p. 306-307.
13. Basch, E., A.M. Deal, M.G. Kris, H.I. Scher, C.A. Hudis, P. Sabbatini, L. Rogak, A.V. Bennett, A.C. Dueck, T.M. Atkinson, J.F. Chou, D. Dulko, L. Sit, A. Barz, P. Novotny, M. Fruscione, J.A. Sloan, and D. Schrag, *Symptom Monitoring With Patient-Reported Outcomes During Routine Cancer Treatment: A Randomized Controlled Trial.* J Clin Oncol, 2016. **34**(6): p. 557-65.
14. Basch, E., L. Barbera, C.L. Kerrigan, and G. Velikova, *Implementation of patient-reported outcomes in routine medical care.* American Society of Clinical Oncology Educational Book, 2018. **38**: p. 122-134.
15. Coon, C.D. and J.C. Cappelleri, *Interpreting Change in Scores on Patient-Reported Outcome Instruments.* Ther Innov Regul Sci, 2016. **50**(1): p. 22-29.

16. Ousmen, A., C. Touraine, N. Deliu, F. Cottone, F. Bonnetain, F. Efficace, A. Bredart, C. Mollevi, and A. Anota, *Distribution- and anchor-based methods to determine the minimally important difference on patient-reported outcome questionnaires in oncology: a structured review.* Health Qual Life Outcomes, 2018. **16**(1): p. 228.
17. Giesinger, J., F. Loth, N. Aaronson, J. Arraras, G. Caocci, F. Efficace, M. Groenvold, M. van Leeuwen, M. Petersen, J. Ramage, K. Tomaszewski, T. Young, and B. Holzner, *Establishing thresholds for clinical importance to improve interpretation of the EORTC QLQ-C30 in daily clinical practice and research.* submitted elsewhere.
18. Petersen, M.A., N.K. Aaronson, J.I. Arraras, W.C. Chie, T. Conroy, A. Costantini, J.M. Giesinger, B. Holzner, M.T. King, S. Singer, G. Velikova, I.M. Verdonck-de Leeuw, T. Young, and M. Groenvold, *The EORTC computer-adaptive tests measuring physical functioning and fatigue exhibited high levels of measurement precision and efficiency.* J Clin Epidemiol, 2013. **66**(3): p. 330-9.
19. Holzner, B., J.M. Giesinger, J. Pinggera, S. Zugal, F. Schopf, A.S. Oberguggenberger, E.M. Gamper, A. Zabernigg, B. Weber, and G. Rumpold, *The computer-based health evaluation system (CHES): a software for electronic patient-reported outcome monitoring.* BMC Med Inform Decis Mak, 2012. **12**(1): p. 126.
20. van de Poll-Franse, L.V., F. Mols, C.M. Gundy, C.L. Creutzberg, R.A. Nout, I.M. Verdonck-de Leeuw, M.J. Taphoorn, and N.K. Aaronson, *Normative data for the EORTC QLQ-C30 and EORTC-sexuality items in the general Dutch population.* Eur J Cancer, 2011. **47**(5): p. 667-75.
21. Giesinger, J.M., N.K. Aaronson, J.I. Arraras, F. Efficace, M. Groenvold, J.M. Kieffer, F.L. Loth, M.A. Petersen, J. Ramage, K.A. Tomaszewski, T. Young, B. Holzner, and E.Q.o.L. Group, *A cross-cultural convergent parallel mixed methods study of what makes a cancer-related symptom or functional health problem clinically important.* Psychooncology, 2018. **27**(2): p. 548-555.
22. Giesinger, J.M., W. Kuijpers, T. Young, K.A. Tomaszewski, E. Friend, A. Zabernigg, B. Holzner, and N.K. Aaronson, *Thresholds for clinical importance for four key domains of the EORTC QLQ-C30: physical functioning, emotional functioning, fatigue and pain.* Health Qual Life Outcomes, 2016. **14**: p. 87.
23. Hosmer, D., S. Lemeshow, and R. Sturdivant, *Applied logistic regression. Third edition.* 2013, Hoboken, New Jersey: John Wiley & Sons.
24. Youden, W.J., *Index for rating diagnostic tests.* Cancer, 1950. **3**(1): p. 32-5.
25. NCSS-LCC, *PASS - Power Analysis and Sample Size*, 2011: Kaysville (US).
26. Cella, D., S. Choi, S. Garcia, K.F. Cook, S. Rosenbloom, J.S. Lai, D.S. Tatum, and R. Gershon, *Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment.* Qual Life Res, 2014. **23**(10): p. 2651-61.
27. Nagaraja, V., C. Mara, P.P. Khanna, R. Namas, A. Young, D.A. Fox, T. Laing, W.J. McCune, C. Dodge, D. Rizzo, M. Almackenzie, and D. Khanna, *Establishing clinical severity for PROMIS((R)) measures in adult patients with rheumatic diseases.* Qual Life Res, 2018. **27**(3): p. 755-764.
28. Cook, K.F., D.E. Victorson, D. Cella, B.D. Schalet, and D. Miller, *Creating meaningful cut-scores for Neuro-QOL measures of fatigue, physical functioning, and sleep disturbance using standard setting with patients and providers.* Qual Life Res, 2015. **24**(3): p. 575-89.
29. Taminiau-Bloem, E.F., F.J. van Zuuren, M.A. Koeneman, B.D. Rapkin, M.R. Visser, C.C. Koning, and M.A. Sprangers, *A 'short walk' is longer before radiotherapy than afterwards: a qualitative study questioning the baseline and follow-up design.* Health Qual Life Outcomes, 2010. **8**: p. 69.
30. Snyder, C., M. Brundage, Y.M. Rivera, and A.W. Wu, *A PRO-cision Medicine Methods Toolkit to Address the Challenges of Personalizing Cancer Care Using Patient-Reported Outcomes: Introduction to the Supplement*, 2019, LWW.

31. Loth, F., B. Holzner, M. Sztankay, H. Bliem, S. Raoufi, G. Rumpold, and J. Giesinger, *Cancer patients' understanding of longitudinal EORTC QLQ-C30 scores presented as bar charts.* Patient education and counseling, 2016. **99**(12): p. 2012-2017.
32. Snyder, C.F., K.C. Smith, E.T. Bantug, E.E. Tolbert, A.L. Blackford, M.D. Brundage, and P.D.P.S.A. Board, *What do these scores mean? Presenting patient-reported outcomes data to patients and clinicians to improve interpretability.* Cancer, 2017. **123**(10): p. 1848-1859.
33. Absolom, K., A. Gibson, and G. Velikova, *Engaging Patients and Clinicians in Online Reporting of Adverse Effects During Chemotherapy for Cancer: The eRAPID System (Electronic Patient Self-Reporting of Adverse Events: Patient Information and aDvice).* Medical care, 2019. **57**: p. S59-S65.
34. Girgis, A., I. Durcinoska, A. Arnold, and G.P. Delaney, *Interpreting and Acting on the PRO Scores From the Patient-reported Outcomes for Personalized Treatment and Care (PROMPT-Care) eHealth System.* Medical care, 2019. **57**: p. S85-S91.

Tables

Table 1: Descriptive statistics for sociodemographic and clinical variables (n=498)

	Mean (SD)	**Range**	
Age	60.4 (12.7)	19-87	
		N	**%**
Sex	Women	272	55.2
	Men	221	44.8
	Missing data	5	
Diagnosis	Breast cancer	117	23.6
	Haematological malignancy	66	13.3
	Lung cancer	49	9.9
	Prostate cancer	48	9.7
	Colorectal cancer	42	8.5
	Head and neck cancer	39	7.9
	Lymphoma	37	7.5
	Gynaecologic cancer	29	5.9
	Stomach cancer	12	2.4
	Brain cancer	10	2.0
	Other	46	9.3
	Missing data	3	
UICC stage*	I	61	16.1
	II	100	26.4
	III	79	20.8
	IV	139	36.7
	Missing data	16	
Comorbidity	No	272	59.0
	Yes	189	41.0
	Missing data	37	
Treatment intention	Curative	282	60.6
	Palliative	183	39.4
	Missing data	33	
Current treatment	No current treatment	115	23.3
	Current treatment	379	76.7
	Surgery**	135	35.6
	Chemotherapy**	232	61.2
	Radiotherapy**	127	33.5
	Other**	81	21.4
	Missing data	4	

*Only reported for patients with solid tumours

**More than one treatment is possible, so the percentages for the treatment modalities do not sum up to 100%

UICC=Union for International Cancer Control

Table 2: Comparison of EORTC CAT Core in patients with clinically important problems/symptoms (cases) and those without (non-cases)

EORTC QLQ-C30 scale	Non-cases			Cases					
	Prevalence	Mean	SD	Prevalence	Mean	SD	Mean difference	Pooled SD	Effect size*
Functioning scales									
Physical Functioning (PF)	45.5%	50.0	7.8	54.5%	38.2	9.4	-11.8	8.7	-1.36
Role Functioning (RF)	75.4%	44.5	8.6	24.6%	30.2	6.4	-14.3	8.1	-1.76
Social Functioning (SF)	81.5%	45.9	7.9	18.5%	35.6	7.0	-10.3	7.7	-1.34
Emotional Functioning (EF)	72.0%	52.0	7.8	28.0%	39.5	6.3	-12.5	7.5	-1.67
Cognitive Functioning (CF)	88.5%	49.1	8.2	11.5%	36.9	8.6	-12.2	8.2	-1.48
Symptom scales									
Fatigue (FA)	64.5%	50.5	7.4	35.5%	64.9	6.6	14.4	7.1	2.02
Pain (PA)	82.0%	47.3	8.1	18.0%	64.2	7.1	16.9	7.9	2.13
Nausea/Vomiting (NV)	89.7%	51.5	9.1	10.3%	78.9	14.9	27.4	9.8	2.79
Dyspnoea (DY)	80.8%	51.2	8.9	19.2%	66.2	5.1	15.0	8.4	1.79
Sleep Disturbances (SL)	81.8%	50.1	8.1	18.2%	63.0	6.7	12.9	7.8	1.65
Appetite Loss (AP)	89.1%	54.1	10.5	10.9%	72.6	6.0	18.5	10.1	1.83
Constipation (CO)	90.1%	49.6	9.5	9.9%	70.0	8.0	20.3	9.4	2.17
Diarrhoea (DI)	91.7%	50.2	9.9	8.3%	72.9	8.1	22.7	9.8	2.32
Financial Impact (FI)	85.8%	48.6	7.5	14.2%	69.0	8.8	20.4	7.7	2.65

*Effect size Cohen's d = Mean difference / Pooled SD

Table 3: Results of the receiver operator characteristic (ROC) analysis and thresholds for clinical importance

QLQ-C30 scale	TCI	Sensitivity	Specificity	AUC	95% CI
Functioning scales					
Physical Functioning (PF)	46	0.82	0.66	0.84	0.80-0.87
Role Functioning (RF)	37	0.84	0.79	0.91	0.88-0.94
Social Functioning (SF)	41	0.80	0.69	0.84	0.79-0.89
Emotional Functioning (EF)	46	0.86	0.71	0.89	0.86-0.93
Cognitive Functioning (CF)	45	0.82	0.67	0.85	0.79-0.90
Symptom scales					
Fatigue (FA)	57	0.92	0.84	0.94	0.91-0.96
Pain (PA)	56	0.90	0.79	0.93	0.90-0.96
Nausea/Vomiting (NV)	58	0.90	0.82	0.92	0.88-0.97
Dyspnoea (DY)	60	0.93	0.77	0.93	0.91-0.95
Sleep Disturbances (SL)	55	0.91	0.76	0.89	0.86-0.93
Appetite Loss (AP)	63	0.94	0.75	0.94	0.91-0.96
Constipation (CO)	57	0.96	0.73	0.94	0.90-0.97
Diarrhoea (DI)	62	0.95	0.82	0.94	0.90-0.98
Financial Impact (FI)	58	0.93	0.83	0.94	0.91-0.97

Please note that for the functioning scales, scoring equal to or below the TCI indicates a clinically important problem, whereas for the symptom scales, scores equal to or above the TCI indicate such a problem.

TCI=Threshold for clinical importance; AUC=Area under curve; CI=Confidence interval

Figures

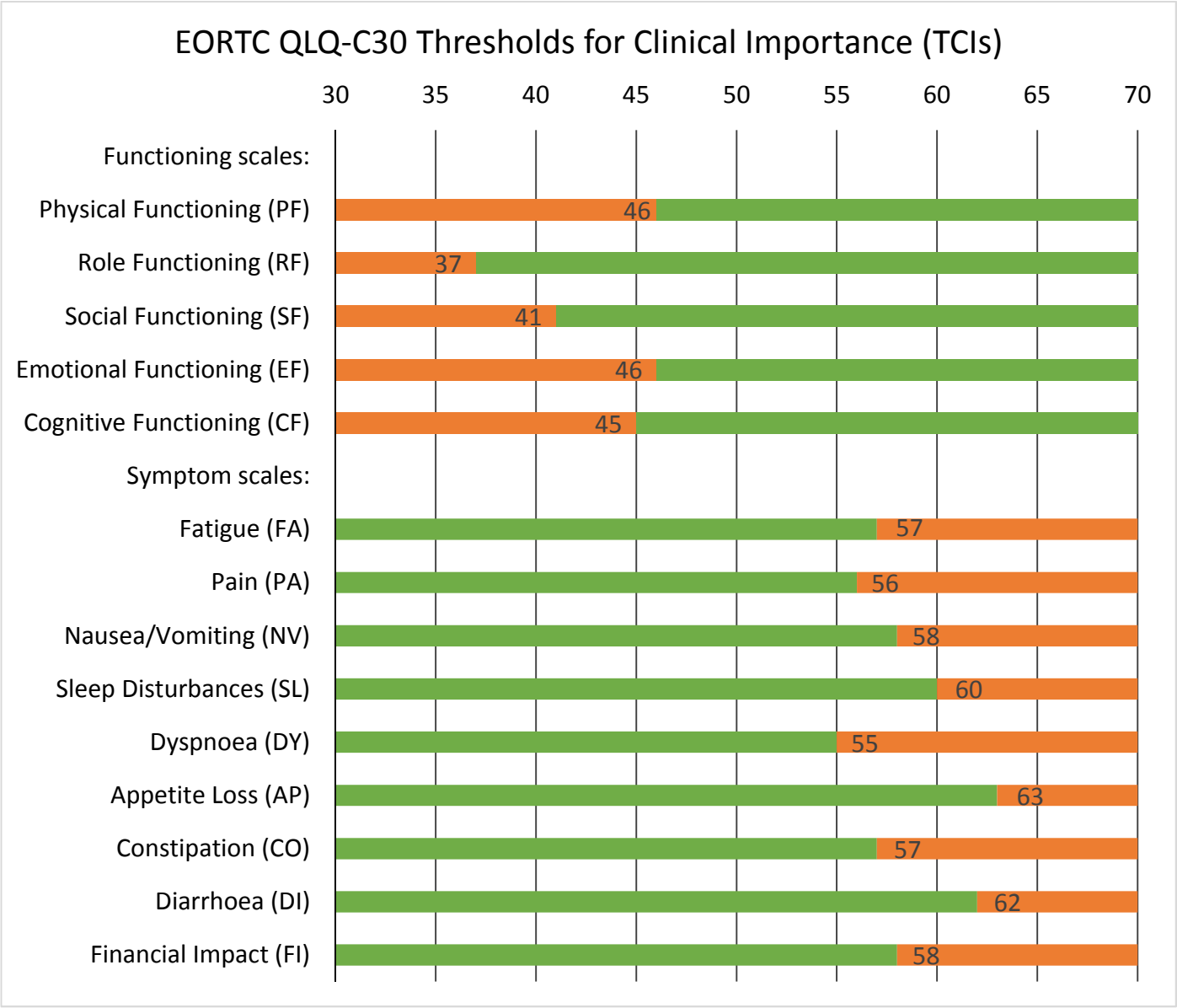

Figure 1: Thresholds for clinical importance (TCIs) for the functioning and symptom scales of the EORTC CAT Core. TCIs are shown inside the bars. Patient scores in the orange range of the bar (i.e. equal or below the TCI for functioning scales, and equal or above the TCI for symptom scales) indicate clinically important problems or symptoms.

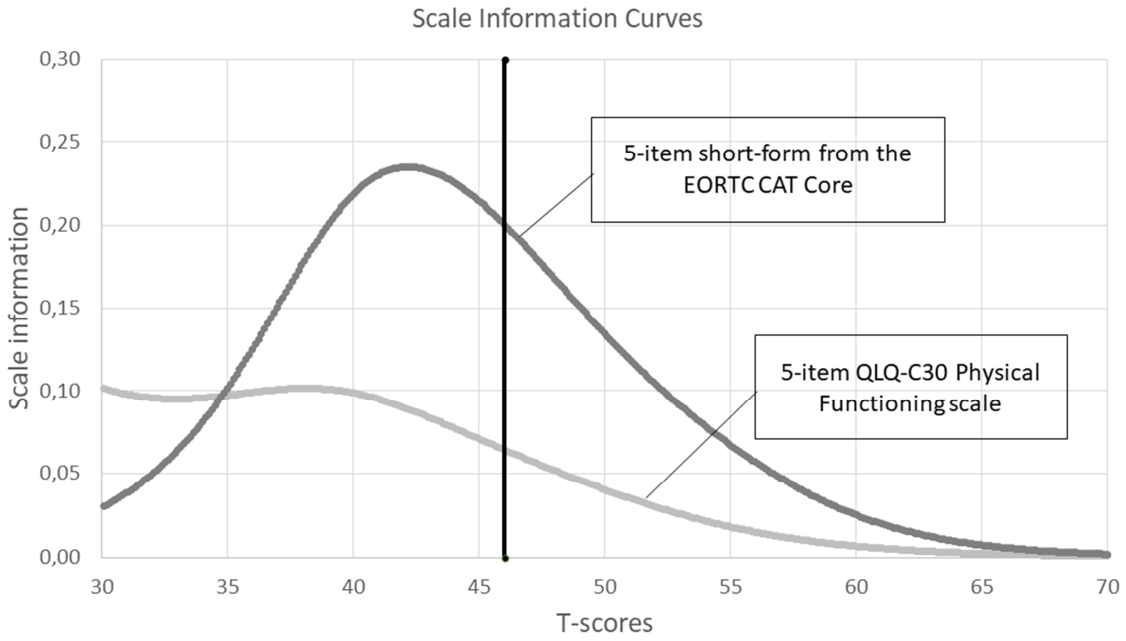

Figure 2: Item information curves showing measurement precision for the 5-item QLQ-C30 Physical Functioning scale, and for a 5-item short-form from the EORTC CAT Core item bank that was created specifically to maximize measurement precision for scores close to the TCI (vertical black line)