Health and Quality of Life Outcomes

RESEARCH Open Access

Using prior information to individualize start item selection when assessing physical functioning with the EORTC CAT Core

Morten Aa. Petersen[1,5*], Hugo Vachon[2], Johannes M. Giesinger[3], Mogens Groenvold[1,4] and on behalf of the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group

Abstract

Background Computerized adaptive test (CAT) provides individualized measurement, using the patient's previous responses to select the next most informative item. However, the first item, the start item, is usually not individualized as no score estimate is available a priori. The European Organisation for Research and Treatment of Cancer (EORTC) CAT Core covers 15 health-related quality of life domains. Scores for one domain may be used to obtain initial score estimates for another domain. We assessed the potential for using such cross-domain information to individualize start item selection for the EORTC CAT Core physical functioning.

Methods The potential for predicting physical functioning (PF) scores from each of the 14 other domains using linear regression was assessed in an international, mixed sample comprising 10,084 cancer patient assessments. Using Monte Carlo CAT simulations, the impact of individually selected PF start items vs. fixed start item for CAT measurement precision was assessed.

Results Depending on the domain predicting PF, the correlation of predicted and observed PF scores ranged 0.25–0.71 and the predicted PF scores were within 1SD of the observed PF scores for 57–85% of the patients. The CAT simulations showed that individually selected start items improved measurement precision for the initial steps of CATs. The application of individual start items had trivial or no impact on measurement precision when the CAT asked three or more items.

Conclusions Simple linear regression may provide useful cross-domain predictions. Using individualized start items may increase measurement precision of the EORTC CAT Core for the initial steps of CAT which may be of relevance for short CATs.

Keywords CAT, EORTC CAT Core, Prior information, Simulation, Start item

*Correspondence:
Morten Aa. Petersen
Morten.Aagaard.Petersen@regionh.dk
[1]Palliative Care Research Unit, Department of Geriatric and Palliative Medicine, Copenhagen University Hospital, Bispebjerg & Frederiksberg, Copenhagen, Denmark
[2]Quality of Life Department, European Organization for Research and Treatment of Cancer, Brussels, Belgium
[3]Department of Psychiatry, Psychotherapy, Psychosomatics, and Medical Psychology, Innsbruck Medical University, Innsbruck, Austria
[4]Department of Public Health, University of Copenhagen, Copenhagen, Denmark
[5]Palliative Care Research Unit, Department of Geriatric and Palliative Medicine, Bispebjerg & Frederiksberg Hospital, Bispebjerg bakke 23B, Copenhagen, NV 2400, Denmark

Background

The European Organisation for Research and Treatment of Cancer (EORTC) has developed an adaptive instrument, the EORTC CAT Core, covering the same health related quality of life (HRQoL) domains as the EORTC QLQ-C30 questionnaire [1–4]. Computerized adaptive tests (CATs) like the EORTC CAT Core tailor the set of presented items to the individual [5, 6]. This has several advantages, including more relevant items and increased measurement precision, compared to traditional, standardised questionnaires where everybody is asked the same set of items [5, 6].

In CAT-assessment, items are selected and presented sequentially. At each step, the CAT uses a score estimate based on responses to the previously asked items to assess which item seems most informative to ask. This individual item selection can be applied as soon as a score estimate is available. However, for the first item, i.e., the start item, the individual's score level is typically unknown. Although there are several approaches for selecting the start item [6, 7], it is often selected a priori, and particularly in HRQoL it is common that the same start item is used for everyone completing a specific CAT (see e.g [8–11]). This approach is also applied for the EORTC CAT Core, i.e., all assessments based on the same CAT-setting (the set of criteria defining the CAT-assessment) are initiated with the same item. However, the same item is rarely optimal for all participants in a study. If prior information about an individual's score level was available, the start item selection could be tailored to the individual thereby potentially increase the measurement precision of EORTC CAT assessments.

Information about an individual's score level available prior to initiating a CAT assessment has been termed 'collateral', 'empirical prior', and 'out-of-scale' information and 'inter-subtest branching' [12–16]. Here we will refer to such information as prior information. We have only identified a limited number of studies investigating the use of prior information in CAT [12, 14–19], and only one within HRQoL measurement [16]. Several of the studies used the prior information both for selecting start item and estimating domain score [15, 16, 19]. Including prior information in the domain score estimation may reduce CAT length and/or increase measurement precision [16]. However, it also means that patients responding to the same items in the same way will get different score estimates if the prior information differs. Here, this could e.g., mean that patients providing the same responses to the same physical functioning items could still get different physical functioning scores if they had different fatigue scores. This deviates from the current scoring of the EORTC CAT Core, and any other EORTC instrument, for which the score of a domain depends solely on the responses to the items of that specific domain. We retained this scoring principle, i.e., prior information was not included in domain score estimation but was based solely on responses to items of the particular domain. Hence, investigations focused on the potential for and impact of using prior information for the selection of start item in assessments with the EORTC CAT Core.

In principle, any prior information providing knowledge about a domain may be used to estimate an individual 'a priori' score. The prior information available will differ across studies and most cannot be obtained within a CAT-assessment but needs to be supplied by external sources. Within the EORTC CAT Core, 14 symptom and functional domains plus overall health/quality of life can be assessed. Thus, when at least two domains are measured, score estimates from other domains are available prior to all but the first domain assessed. That is, in a multi-domain assessment, estimates for other domains may be used as prior information to select start item and this prior information can be obtained entirely within the CAT-assessment. To be of practical relevance, it must be possible to predict scores with 'reasonable' precision, and this would be particularly useful if it could be done in a simple and efficient way.

The aim of this study was two-fold: (1) to evaluate whether scores on one domain can be predicted from scores on another domain, and (2) to assess whether individually selected start items may improve measurement precision with the EORTC CAT Core.

This study explores a novel approach to start item selection in the context of HRQoL CAT assessment. Hence, the study should be viewed as a proof-of-concept, exploring the feasibility and potential advantages of this approach within the specific framework of the EORTC CAT Core. While the primary focus is on this instrument, our findings may provide a foundation for future research on start item selection in other CAT systems.

Methods

The EORTC CAT core

The EORTC CAT Core is a CAT system including item banks for the five functional and nine symptom domains of the QLQ-C30 questionnaire [3] plus the two QLQ-C30 items on overall health/quality of life [20]. The EORTC CAT Core and its customised software allows for flexible assessment of these domains, i.e., users of the system may design CATs matching their needs, including selecting which domains to assess, how to select items, and how many items to ask (for more detail on the use of the EORTC CAT Core, visit the EORTC QLG website https://qol.eortc.org/questionnaires/core/cat/). Each of the item banks comprises between 7 and 34 items, with a total of 260 items, and includes the items of the QLQ-C30. All items of the 14 banks apply the response options: 'not at

all', 'a little', 'quite a bit', and 'very much'. The EORTC CAT Core measures are scored on T-score metrics, scaled so that the European general population has a mean of 50 and a standard deviation of 10 for all domains [21]. This means that for functional domains like physical functioning scores > 50 reflect better functioning than the average of the European general population while for symptom domains (e.g., fatigue) scores > 50 reflect worse symptoms than the average of the European general population. Similarly, scores < 50 reflect poorer functioning/less symptoms than for the European general population average.

We focused on the selection of start item for the physical functioning domain and used this to pilot test whether 'cross-domain' predicted start items may be a viable way of improving measurement. Physical functioning is a key aspect of HRQoL [22] and is generally assessed with high precision in our sample (see below). The physical functioning item bank consists of 31 items of which five originate from the QLQ-C30. A plot of the physical functioning item bank information function is shown in Fig. 1.

Domain score prediction

Sample

For the evaluations of cross-domain prediction we combined the empirical data collected for the development and validation of the EORTC CAT Core [3, 4]. These datasets were international, mixed samples comprising a total of 10,084 cancer patient assessments with T-score estimates of all 15 domains [3, 4, 21]. Depending on the study they had participated in, patients had answered different subsets of the total 260 items and hence, scores were based on different subsets of items (full item bank, QLQ-C30 items only, or CAT assessment). Specifically, physical functioning had been assessed with between five and 31 items. To reduce the risk of overestimating the performance of the prediction models, we split the sample in a training dataset for fitting the prediction models and a testing dataset for evaluating the prediction. We randomly allocated 80% of the data to the training set and 20% to the testing set [23].

Prediction of domain score

When using expected a posteriori (EAP) estimation of the domain score, it is standard to assume that the scores a priori follow a normal distribution [24, 25]. EAP is the standard scoring for the EORTC CAT Core and also the one applied in this study to obtain the T-score estimates. Linear regression, which assumes normally distributed residuals, is a straightforward approach for predicting a continuous outcome. Regressing the domain score on a set of predictors (e.g., other HRQoL scores or patient characteristics) was also suggested by Van der Linden as a simple approach to obtain an initial domain score estimate [14]. When a score for domain X has been estimated, a score for physical functioning (PF) may be predicted using the simple Eq.

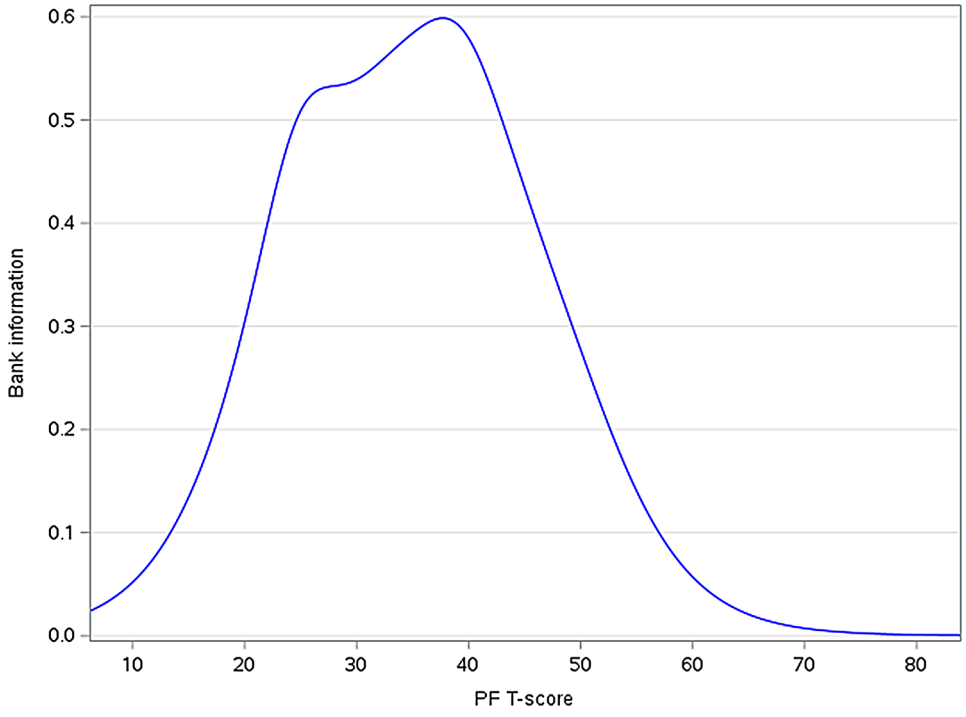

Fig. 1 Bank information function for the physical functioning (PF) item bank

$$PF = \alpha + \beta * X \quad (1)$$

Where α and β are regression parameters which have been estimated beforehand (in the training set). The regression prediction may easily be extended to include two or more domains as predictors, however, here we aimed for a simple prediction that could be used as soon as one EORTC CAT Core domain had been assessed, therefore, we focused on predicting PF by one domain at a time.

Domain scores were estimated based on all available items for each individual. For each of the 14 other domains the regression model (1) was fitted in the training dataset to obtain estimates of α and β. These models were used to produce 14 predicted PF scores for each individual in the testing dataset.

Evaluations of predicted PF scores

For each patient in the testing dataset the differences between the 'observed' PF score estimated based on the available PF items and the PF score predicted from each of the other domains, respectively, were calculated. The mean difference and correlation between predicted and observed PF scores, and the percent predictions deviating < 5 points (=½SD) and deviating < 10 points (= 1SD), respectively, for the 14 cross-domain predictions were estimated and plotted.

Impact of individually selected start items

Simulation sample

To evaluate the impact of individually selected start items versus fixed start item, we conducted a series of Monte Carlo CAT simulations. When a specific population distribution is simulated, the choice of distribution will affect findings and conclusion. As no population was of specific focus but we wanted to assess the impact of individually selected start items generally, we simulated uniformly across the continuum of possible PF scores. The original calibration sample had mean = 47 and had scores within 47 ± 30. To cover this range sufficiently, 200 sets of responses to the 31 PF items were simulated for each PF score in the range 17–77 with increments of 0.5 (17, 17.5,…). That is, for each sampled PF score the probability of the four response options 'not at all', 'a little', 'quite a bit', and 'very much' to each item were estimated using the established item parameters for the generalized partial credit model calibrated for the PF item bank [26, 27]. Based on these item response probabilities, a random item response was selected. From these responses, fixed length CAT-assessments asking 1 to 10 PF items, respectively, were simulated. In each step of the CATs, the PF score was estimated using EAP based on the items asked so far and the item providing the maximum Fisher information in this PF score estimate was selected for the next step [6, 28]. Four different types of start items of the CATs were simulated. One used a fixed start item where all simulated CATs were initiated with the same item: the one most informative at the a priori mean of 47 ('fixed'). Three initiations used individually selected start items each reflecting different levels of predictive power. The first started the CATs with the item most informative at the individual's 'true' PF score, i.e., at the sampled PF score ('true'). This reflects the ideal situation where the PF score is predicted perfectly from another domain. The second started the CATs with the item most informative five points =½SD from the individual's true PF score– half the sample five points above and the other half five below ('diff ½SD'). The last started the CATs with the item most informative 10 points (1SD) from the individual's true PF score ('diff 1SD'). The last two initiations reflect situations where the predicted score deviates from the true score.

Evaluations of individually selected start items

Since the fixed start item was the most informative item at the a priori mean we anticipated that the 'fixed' CATs would provide efficient and precise assessment near the a priori mean, while for more extreme scores we expected a larger advantage of individual start items. Therefore, we divided the evaluations up in three sections: Low PF scores (3SD to 1SD below the mean), middle PF scores (1SD below to 1SD above the mean), and high PF scores (1SD to 3SD above the mean). Findings for each of the three sections may be useful when assessing the effect of start item for a specific population. For example, if a population with predominantly low PF scores is studied, findings for the 'low PF scores' section are likely of most relevance.

For each of the score sections we calculated and plotted the following:

- The mean difference between the true PF score and scores based on CATs with each of the four types of start item.
- The percentage differences < 5 points (½SD) between true and CAT estimated PF scores.
- The mean reliability obtained in each step of CATs with the four types of start item. The reliability in step *k* of the CAT was estimated as $1\text{-}SE(PF)^2/SD^2$ where *SE(PF)* is the standard error of the current PF score estimate, estimated from the Fisher information function based on the *k* items asked by $SE(PF)^2 = 1/information(PF)$ [29]

All analyses and CAT simulations were conducted using SAS Enterprise Guide software version 7.15.

Results

Domain score prediction

Characteristics of the $N = 10,084$ patients in the total sample are shown in Table 1. The sample was a mixed collection of cancer patients representing 12 countries and a variety of cancer diagnoses and treatments. The training dataset included a random subsample of $N = 8,068$ and the testing set included the remaining $N = 2,016$.

Figure 2 shows that all mean differences were < 1 (0.1SD), i.e., PF scores predicted with any of the other domains were on average close to the observed PF scores. Correlations between observed and predicted PF scores varied more, ranging 0.25 (using diarrhoea) to 0.71 (role functioning) with most correlations being of at least moderate size (> 0.4) (Fig. 3). Figure 4 also shows variation across domains in the percent predicted PF scores deviating <½SD and < 1SD, respectively, from the observed score. Percent scores deviating <½SD ranged from 33% (cognitive functioning) to 50% (fatigue) while deviations < 1SD ranged from 57% (diarrhoea) to 85% (role functioning).

Table 1 Clinical and sociodemographic characteristics of the $N = 10,084$ patients in the sample

		N	percentage
Age, mean year(SD)		10,025	60.4(13.1)
	Missing	59	
Sex	Female	5,437	54.2%
	Male	4,599	45.8%
	Missing	48	
Country	Australia	114	1.1%
	Austria	374	3.7%
	Denmark	2,043	20.3%
	France	1,010	10.0%
	Germany	300	3.0%
	Italy	397	3.9%
	Poland	824	8.2%
	Spain	407	4.0%
	Sweden	282	2.8%
	Taiwan	403	4.0%
	The Netherlands	129	1.3%
	UK	3,801	37.7%
	Missing	0	
Cancer stage	I-II	4,676	50.3%
	III-IV	4,616	49.7%
	Missing	792	
Cancer site	Breast	2,088	21.4%
	Gastrointestinal	1,370	14.1%
	Gynaecological	1,164	12.0%
	Head and neck	1,097	11.3%
	Lung	577	5.9%
	Urogenital	1,523	15.6%
	Other	1,911	19.6%
	Missing	347	
Current treatment	Chemotherapy	4,268	43.1%
	Other treatment	2,016	20.4%
	No current treatment	3,612	36.5%
	Missing	188	
Employment status	Retired	4,788	48.4%
	Working	3,500	35.4%
	Other	1,601	16.2%
	Missing	195	
Cohabitation	Living with a partner	7,368	74.5%
	Living alone	2,517	25.5%
	Missing	199	

Impact of individually selected start item

Figure 5A-C summarizes the simulated effect of using the same start item for all and individually selected start items for CATs asking 1 to 10 items, respectively. Generally, the effect of the type of start item tapered off after a few items. For CATs asking four or more items, there was almost no differences between different types of start of item, and often differences were trivial already after asking two items.

For high PF scores (1SD-3SD above the mean) the only non-trivial differences across start items were for the reliability when asking one item. Using the same start item for all resulted in 0.10–0.15 lower reliability than using an individually selected start item. Except for this, using individually selected start items did not seem to affect estimation of high PF scores. The generally lower agreement between CAT and true scores observed for high PF than low and middle PF scores was likely due to a ceiling effect as the PF item bank includes few items particularly informative for those with very good PF.

For middle PF scores (mean ± 1SD) there were only trivial differences across start item types when two or more items were asked. For the first item, the 'fixed', 'true' and 'diff ½SD' CATs resulted in similar performances while starting with the item most informative 1SD from the true score ('diff 1SD') resulted in estimated scores deviating slightly more from the true score.

For low PF scores (1SD-3SD below mean) using the same start item for all resulted in larger deviations from the true score for the first item both at group level (mean deviations 2-3.5 points larger) and at the individual level (25–30% fewer estimated scores within ½SD of the true score) than when individually selected start items were used. Asking two or more items there were only trivial variation across start item types in the mean score differences and percent within ½SD of the true score. The most pronounced difference was observed for the reliability of score estimates. For the first item, the CATs with individually selected start items provided 0.48–0.68 higher reliability (increase from 0.21 to 0.69–0.89). Asking two items the individual start items increased reliability with 0.07–0.13 (0.79 to 0.86–0.92) and asking three items 0.03–0.05 higher reliability was observed (0.89 to

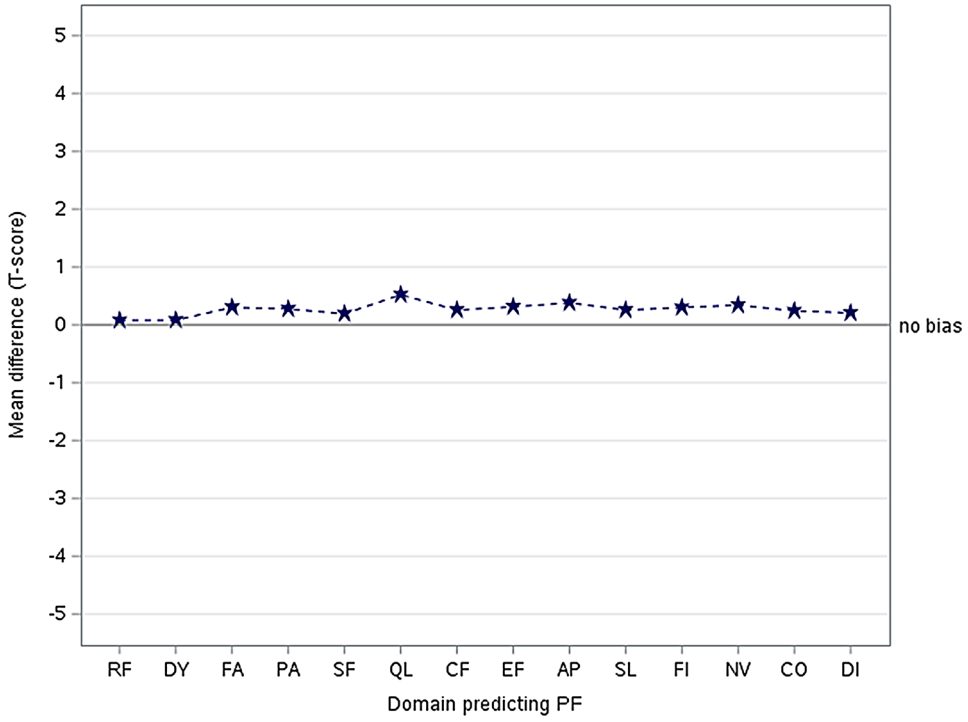

Fig. 2 Mean differences between predicted and observed physical functioning (PF) scores. *RF* role functioning; *DY* dyspnoea; *FA* fatigue; *PA* pain; *SF* social functioning; *QL* overall health/quality of life; *CF* Cognitive functioning; *EF* emotional functioning; *AP* lack of appetite; *SL* insomnia; *FI* financial difficulties; *NV* nausea & vomiting; *CO* constipation; *DI* diarrhoea

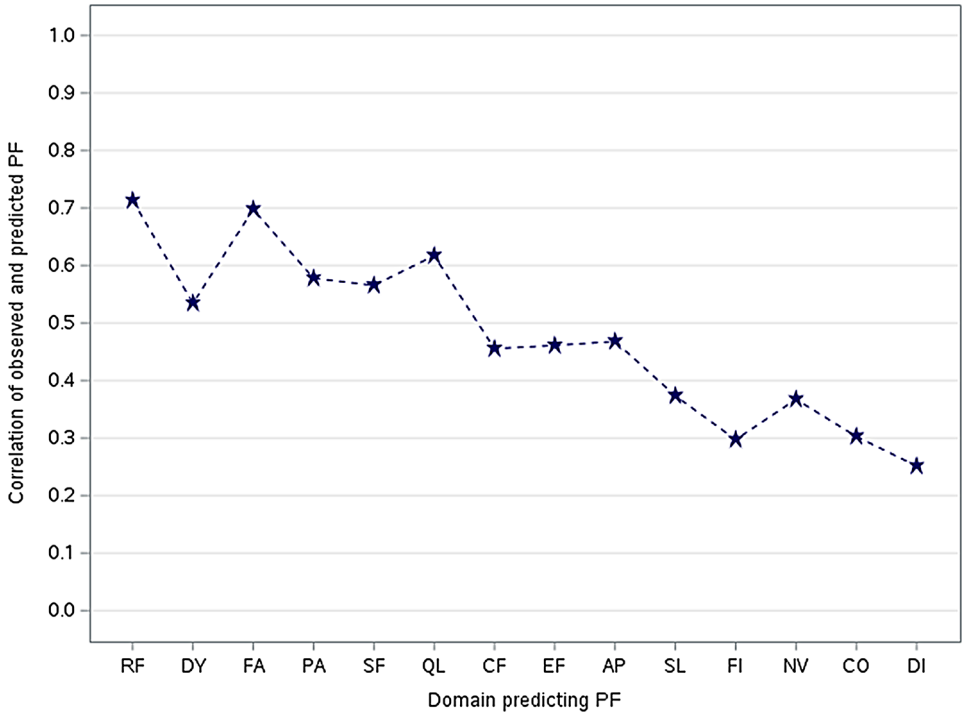

Fig. 3 Correlations between predicted and observed physical functioning (PF) scores. *RF* role functioning; *DY* dyspnoea; *FA* fatigue; *PA* pain; *SF* social functioning; *QL* overall health/quality of life; *CF* Cognitive functioning; *EF* emotional functioning; *AP* lack of appetite; *SL* insomnia; *FI* financial difficulties; *NV* nausea & vomiting; *CO* constipation; *DI* diarrhoea

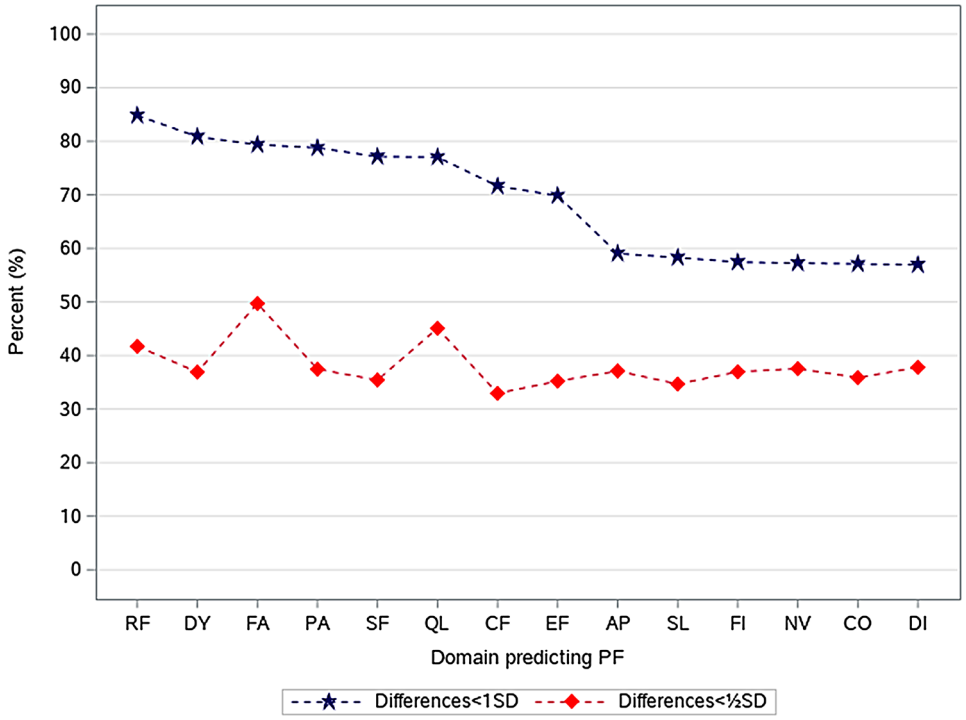

Fig. 4 Percent differences < 1SD and <½SD, respectively, between predicted and observed physical functioning (PF) scores. *RF* role functioning; *DY* dyspnoea; *FA* fatigue; *PA* pain; *SF* social functioning; *QL* overall health/quality of life; *CF* Cognitive functioning; *EF* emotional functioning; *AP* lack of appetite; *SL* insomnia; *FI* financial difficulties; *NV* nausea & vomiting; *CO* constipation; *DI* diarrhoea

0.92–0.94). When asking four or more items the differences in reliability were < 0.02.

Discussion

CAT-assessment adapts the list of asked items to the individual. In each step of a CAT-assessment this adaptation is obtained by using the current score estimate to predict the most relevant next item to ask the respondent. However, in typical CAT-assessments the start item is not individualized as individual score estimates are not available a priori. We investigated whether scores on the EORTC CAT Core physical functioning (PF) could be predicted from the other domains covered by the instrument and whether using individually selected start items could improve measurement precision of PF CAT-assessment. In short, PF scores could be predicted with some variability at the individual level and individually selected start items could improve measurement precision, but the impact only seemed of practical relevance for the first step or two of a CAT-assessment.

As expected, the ability to predict PF scores differed across domains. For eight of the 14 domains, > 70% of the predicted scores were within 1SD of the observed PF score. This implies that in most cases the predicted score can be expected to be reasonably close to the true score and hence, that a start item selected based on a cross-domain predicted score often will be relevant. Note that the prediction accuracy may rely on the precision of the domain score estimate used to predict PF - lower precision could result in lower accuracy. As our evaluations were conducted on a dataset with score estimates of varying precision our findings may be viewed as an estimate of the 'average' performance of cross-domain prediction of PF. The findings have two other implications. First, the order of which domains are assessed matters. For instance, if PF is the primary outcome, one should measure at least one 'good' predictor like role functioning, dyspnoea, or fatigue, before measuring PF. That is, the assessment order of domains should be selected carefully. In general, investigating the 'optimal' sequence of domains for different settings may be an area for future research. Second, for some individuals the predicted score may deviate from the true score to a degree that a start item of limited relevance is selected. When using the same start item for all, the start item will often also be of limited information for some. This is particularly evident from Fig. 5c which shows that using the same start item for all (the most informative item at the sample mean) resulted in an average reliability of 0.2 for those with low PF scores, while choosing an individual start item resulted in an average reliability of at least 0.7 even when choosing the item most informative 1SD from the individual's actual PF score. Note that the relatively low reliabilities observed for high PF scores (see Fig. 5c) reflect a likely ceiling effect in the sense that the item bank may lack informative items for very high PF scores. However,

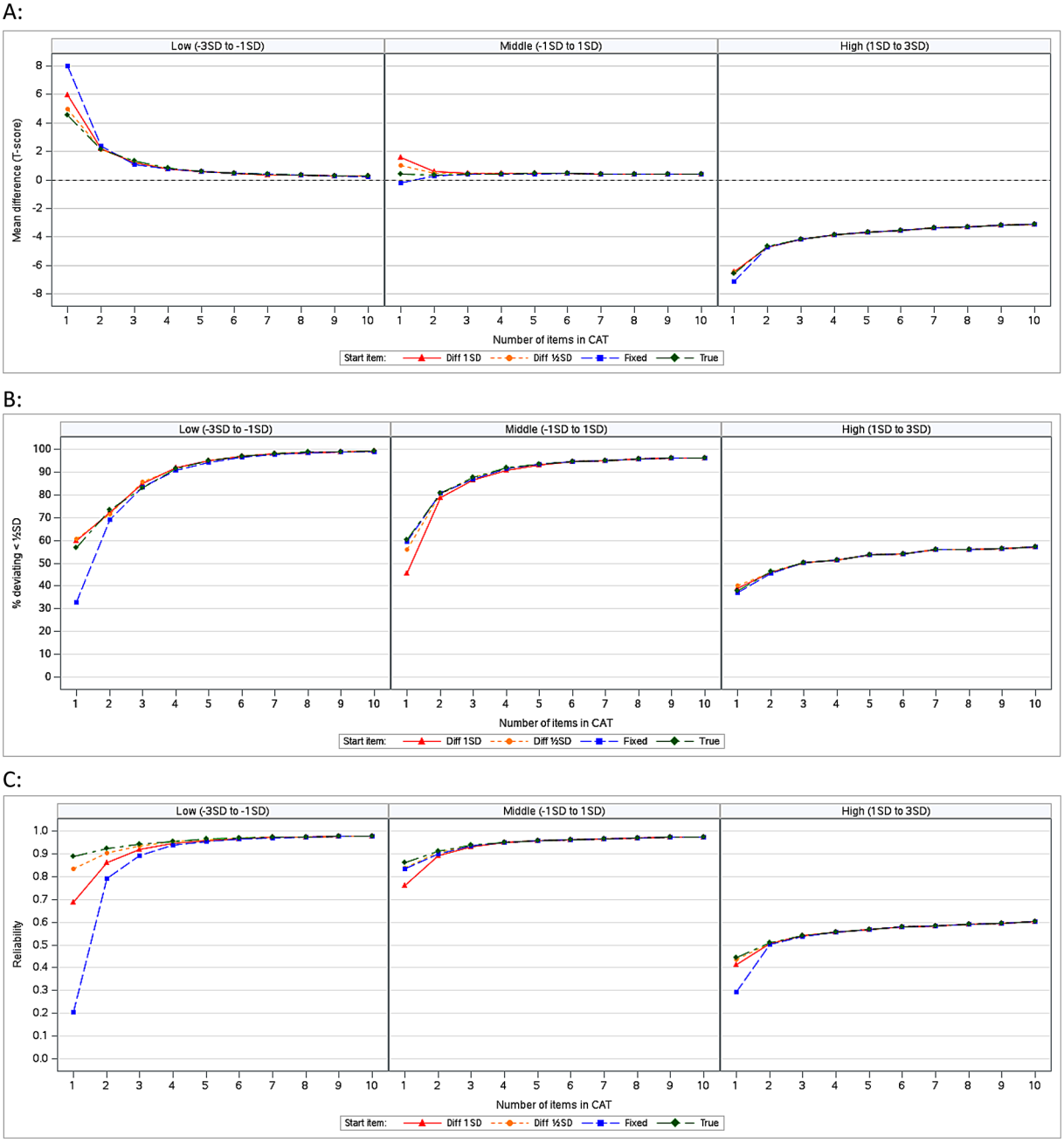

Fig. 5 Mean differences (**A**) and percentage differences <½SD (**B**) between true and CAT estimated physical functioning (PF) scores and mean reliability (**C**) with each type of start item. *Diff 1SD* CATs with individual start item with maximum information 1SD from true PF score; *Diff ½SD* Individual start item with maximum information ½SD from true PF score; *Fixed* fixed start item, i.e., starting all CATs with the same item; *True* Individual start item with maximum information at true PF score

validation of the EORTC CAT Core found a ceiling effect of less than 1% in a mixed sample of cancer patients, i.e., ceiling effect is likely a rare problem in real-life use of the PF item bank [4].

The CAT simulations indicated that using individualized start items only impacted measurement precision of the initial steps of CAT assessments. Few differences were observed when asking three or more items, i.e., individualized start items may have limited impact on measurement precision for CATs asking at least three items. This finding reflects the ability of the CAT procedure to 'track down' within a few items the physical functioning level of the respondent regardless of the starting point. This is reassuring for the cases where a CAT is started with a poorly chosen item. Importantly, the simulations did not indicate that using individualized

start items results in poorer CAT assessments. That is, it seems a viable alternative to the standard of using a fixed start item, with particular relevance for short CATs. For example, when assessing several domains, it may be judged that only a limited number of items can be asked per domain to avoid overburdening patients. With the increased precision of individualized start items, asking 1–2 items per domain may in some cases be enough, e.g., when screening for symptoms/problems (although this should be verified in the specific applications). This may be of particular interest in a busy, clinical setting where the interest is on getting a comprehensive picture of possible symptoms/problems across multiple domains as efficiently as possible. With such an approach, an assessment of the 14 domains of the EORTC CAT Core may be accomplished within a few minutes by most patients [4]. In addition to improved efficiency, well-selected start items increase the chance that patients received relevant and meaningful questions, even when only a few items per domain are asked. This could improve the patient experience, and therefore, involve potential benefits on patients' perceived burden and compliance.

Clear strengths of the study are the large sample and that both observed and simulated data was applied. Cross-domain prediction of PF was evaluated in a large, mixed, international sample of cancer patients split in a training and testing dataset for maximum generalizability to assessment in cancer patients in general. A limitation of the observed sample is that the majority of domain scores were estimated based on QLQ-C30 items only resulting in lower score precision than if the full item banks had been available. This may have resulted in a lower estimated ability to predict scores, i.e., with more precise score estimates even better prediction may be obtained. Still, the scores are on the same metric, i.e., fully compatible with any scores from the item banks. Since a CAT can be set up in numerous ways clearly not all can be evaluated. In the simulations, we compared the impact of individual start items to starting with the most informative item at the centre of the score range for fixed length CATs asking 1, 2,…, 10 items, respectively. Other CAT-settings may result in different findings. It seems plausible e.g., that starting with a more 'extreme' item (e.g., one relevant for high PF scores) would favour the use of individualized start items even more for those at the opposite end of the score range (low PF).

We used simple linear regression to predict PF scores from other domains. This has also been suggested previously as a straightforward approach to obtain an initial domain score estimate [14]. The simple linear regression was informative for our purposes, but alternative prediction models are clearly possible and exploring more sophisticated approaches to obtain initial score estimates may be an area of future research. We used prior information to select start item. Prior information may also be incorporated in the domain score estimation during a CAT assessment. However, this requires a more complex approach, including the use of a generalisation of standard IRT models to also obtain estimation of the uncertainty/variance [14].

This study focuses specifically on the EORTC CAT Core and evaluates the proposed start item selection method within this system. While our findings suggest that simple linear regression can provide useful score prediction to guide individualization of start item selection, it is important to note that the results are limited to the EORTC CAT Core. The performance of this approach in other CAT systems, particularly those assessing different domains, having different item bank characteristics, or employing alternative item selection strategies, remains an open question and warrants further investigation. Such research would provide valuable insights into the generalisability of the method to other domains and CAT systems. Despite being limited to the EORTC CAT Core, this study serves as an important proof-of-concept, demonstrating how predictive modelling may be used to individualise start item selection and thereby enhance measurement precision in the early stages of CAT assessments. Furthermore, the simplicity of the proposed method allows it to be implemented in most multi-domain CAT systems, making it broadly applicable.

Conclusions

The study indicates that simple linear regression may provide useful cross-domain predictions of EORTC CAT Core physical functioning scores which may be used to select more informative start items. Using individualized start items may increase measurement precision for the initial steps of CAT but after a few items the impact seems trivial. Before implementing cross-domain based individualized start items for the EORTC CAT Core, the performance of cross-domain predictions should be investigated for the other domains covered by the EORTC CAT Core.

Abbreviations

AP	Lack of appetite
CAT	Computerized adaptive test
CF	Cognitive functioning
CO	Constipation
DI	Diarrhoea
DY	Dyspnoea
EAP	Expected a posteriori
EF	Emotional functioning
EORTC	European Organisation for Research and Treatment of Cancer
FA	Fatigue
FI	Financial difficulties
HRQoL	Health related quality of life
NV	Nausea & vomiting
PA	Pain
QL	Overall health/quality of life
RF	Role functioning

SF Social functioning
SL Insomnia

Acknowledgements
The authors would like to thank the participating patients as well as the EORTC Quality of Life Group for their essential input and evaluation.

Author contributions
MAP conducted the analyses and wrote the first draft of the manuscript. MAP and MG designed and developed the study. All authors contributed to the interpretation of results, manuscript preparation and approved the final version of the manuscript.

Funding
Open access funding provided by Copenhagen University. The work conducted by MAP was funded by a research grant from the EORTC Quality of Life Group (QLG grant no. 003/2017).

Data availability
The data are not publicly available.

Declarations

Ethics approval and consent to participate
Local ethics committees of the participating countries approved data collection and written informed consent was obtained before study participation. All procedures performed involving human participants were in accordance with the local ethical standards and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Competing interests
The authors declare no competing interests.

Received: 23 January 2025 / Accepted: 27 February 2025
Published online: 09 March 2025

References

1. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, de Haes JC. The European organization for research and treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst. 1993;85(5):365–76.
2. Fayers P, Bottomley A, on behalf of the EORTC Quality of Life Group and of the Quality of Life Unit. Quality of life research within the EORTC - the EORTC QLQ - C30. Eur J Cancer. 2002;38(Suppl 4):S125–33.
3. Petersen MA, Aaronson NK, Arraras JI, Chie W-C, Conroy T, Costantini A, Dirven L, Fayers PM, Gamper EM, Giesinger JM, et al. The EORTC CAT Core - The computer adaptive version of the EORTC QLQ-C30 questionnaire. Eur J Cancer. 2018;100:8–16.
4. Petersen MA, Aaronson NK, Conroy T, Costantini A, Giesinger JM, Hammerlid E, Holzner B, Johnson CD, Kieffer JM, van Leeuwen M, et al. International validation of the EORTC CAT Core– a new adaptive instrument for measuring core quality of life domains in cancer. Qual Life Res. 2020;29(5):1405–17.
5. van der Linden WJ, Glas CAW. Elements of adaptive testing. New York: Springer; 2010.
6. Wainer H. Computerized adaptive testing: A primer. 2nd ed. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.; 2000.
7. Thompson NA, Weiss DJ. A framework for the development of computerized adaptive tests. Practical Assess Res Evaluation. 2011;16(1):1–9.
8. Aletaha D. From the item to the outcome: the promising prospects of PROMIS. Volume 12. Springer; 2010. pp. 1–2.
9. Bjorner JB, Chang CH, Thissen D, Reeve BB. Developing tailored instruments: item banking and computerized adaptive assessment. Qual Life Res. 2007;16(Suppl 1):95–108.
10. Petersen MA, Aaronson NK, Conroy T, Costantini A, Giesinger JM, Hammerlid E, et al. International validation of the EORTC QLQ-CAT– a new adaptive instrument for measuring core quality of life domains in cancer. Qual Life Res. 2020;29(5):1405–17.
11. van der Willik EM, van Breda F, van Jaarsveld BC, van de Putte M, Jetten IW, Dekker FW, Meuleman Y, van Ittersum FJ, Terwee CB. Validity and reliability of the Patient-Reported outcomes measurement information system (PROMIS®) using computerized adaptive testing in patients with advanced chronic kidney disease. Nephrol Dialysis Transplantation. 2023;38(5):1158–69.
12. Brown JM, Weiss DJ. An adaptive testing strategy for achievement test batteries. MINNESOTA UNIV MINNEAPOLIS DEPT OF PSYCHOLOGY; 1977.
13. Kahraman N, Kamata A. Increasing the precision of subscale scores by using Out-of-Scale information. Appl Psychol Meas. 2004;28(6):407–26.
14. van der Linden WJ. Empirical initialization of the trait estimator in adaptive testing. Appl Psychol Meas. 1999;23(1):21–9.
15. Matteucci M, Veldkamp BP. On the use of MCMC computerized adaptive testing with empirical prior information to improve efficiency. Stat Methods Appl. 2013;22(2):243–67.
16. Frans N, Braeken J, Veldkamp BP, Paap MC. Empirical priors in polytomous computerized adaptive tests: risks and rewards in clinical settings. Appl Psychol Meas. 2023;47(1):48–63.
17. Gialluca KA, Weiss DJ. Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement. In Research report 79– 6. University of Minnesota, Department of Psychology, Psychometric Methods; 1979
18. Maurelli VA, Weiss DJ. Factors Influencing the Psychometric Characteristics of an Adaptive Testing Strategy for Test Batteries. 1981.
19. Xie Q. The impact of collateral information on ability Estimation in an adaptive test battery. The University of Iowa; 2019.
20. Petersen MA, Groenvold M, Aaronson NK, Chie W-C, Conroy T, Costantini A, Fayers P, Helbostad J, Holzner B, Kaasa S, et al. Development of computerised adaptive testing (CAT) for the EORTC QLQ-C30 dimensions–General approach and initial results for physical functioning. Eur J Cancer. 2010;46:1352–8.
21. Liegl G, Petersen MA, Groenvold M, Aaronson NK, Costantini A, Fayers PM, Holzner B, Johnson C, Kemmler G, Tomaszewski KA, et al. Establishing the European norm for the health-related quality of life domains of the computer-adaptive test EORTC CAT core. Eur J Cancer. 2019;107:133–41.
22. US Food and Drug Administration, F.D.A. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims. Fed Regist.; 2009.
23. Gholamy A, Kreinovich V, Kosheleva O. Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation. In Departmental Technical Reports (CS), vol. 1209; 2018.
24. Bock RD, Aitkin M. Marginal maximum likelihood Estimation of item parameters: application of an EM algorithm. Psychometrika. 1981;46(4):443–59.
25. Bock RD, Mislevy RJ. Adaptive EAP Estimation of ability in a microcomputer environment. Appl Psychol Meas. 1982;6(4):431–44.
26. Petersen MA, Groenvold M, Aaronson NK, Chie W-C, Conroy T, Costantini A, Fayers P, Helbostad J, Holzner B, Kaasa S, et al. Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 physical functioning dimension. Qual Life Res. 2011;20(4):479–90.
27. Muraki E. A Generalized Partial Credit Model. In Handbook of Modern Item Response Theory. Edited by van der Linden WJ, Hambleton RK. Berlin: Springer; 1997: 153– 68.
28. Muraki E. Information functions of the generalized partial credit model. Appl Psychol Meas. 1993;17(4):351–63.
29. Kim S, Feldt LS. The Estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. Asia Pac Educ Rev. 2010;11(2):179–88.

Publisher's note