Quality of Life Research (2019) 28:3355–3362
https://doi.org/10.1007/s11136-019-02261-2

Establishing clinically-relevant terms and severity thresholds for Patient-Reported Outcomes Measurement Information System® (PROMIS®) measures of physical function, cognitive function, and sleep disturbance in people with cancer using standard setting

Nan E. Rothrock[1] · Karon F. Cook[1] · Mary O'Connor[1] · David Cella[1] · Ashley Wilder Smith[2] · Susan E. Yount[1]

Accepted: 30 July 2019 / Published online: 13 August 2019
© Springer Nature Switzerland AG 2019

Abstract

Purpose Patient-Reported Outcomes Measurement Information System® (PROMIS®) physical function, cognitive function, and sleep disturbance measures are increasingly used in cancer care. However, there is limited guidance for interpreting the clinical meaning of scores. This study aimed to apply bookmarking, a standard setting methodology, to identify PROMIS score thresholds in the context of cancer care.

Methods Using item parameters, we constructed vignettes of five items covering the range of possible scores. Focus groups were held with cancer care providers and people with cancer. Terminology for categorizing levels of severity was explored. Participants rank ordered vignettes by severity and then placed bookmarks between vignettes representing different levels of severity. Group discussion was held until consensus on bookmark placement was reached.

Results Clinicians selected "within normal limits," "mild," "moderate," and "severe" to describe levels of severity. Both patients and clinicians were able to apply these labels, but there was not unanimous support for any set of descriptors. Clinicians and patients agreed on all severity thresholds for sleep disturbance. For cognitive and physical function, clinicians and patients agreed on the threshold between "within normal limits" and "mild." However, patients required greater dysfunction than clinicians before applying "moderate" and "severe" labels.

Conclusions Bookmarking can be applied to develop provisional score interpretation for PROMIS measures. Patients and clinicians were frequently consistent in their bookmark placement. When there was variance, patients required more dysfunction before assigning more severity. Additional research with other cancer samples is needed to evaluate the replicability and generalizability of our findings.

Keywords Patient-reported outcomes · Physical function · Sleep · Cognitive function · Reference values · PROMIS

Introduction

The impact of cancer and its treatment on individuals' function and symptoms is often captured using patient-reported outcome (PRO) measures. For example, fatigue, pain, and emotional distress have been shown to be important outcomes in observational studies and trials, and are increasingly captured in clinical practice [1–4]. Although scores from PRO measures have demonstrated reliability and validity in clinical research, there is limited guidance for identifying score thresholds that communicate the severity level or clinical meaningfulness of a score. Score interpretability is one of 8 scientific criteria established by the Medical

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-019-02261-2) contains supplementary material, which is available to authorized users.

✉ Nan E. Rothrock
n-rothrock@northwestern.edu

[1] Department of Medical Social Sciences, Feinberg School of Medicine of Northwestern University, Chicago, IL, USA

[2] Outcomes Research Branch, National Cancer Institute, Bethesda, MD, USA

Outcomes Trust Scientific Advisory Board to promote "high-quality, standardized, health outcomes measurement instruments to national and international health communities" (p. 979) [5]. Interpretability has long been neglected relative to other psychometric properties such as validity, responsiveness, and reliability. But, with increased use of PROs in clinical settings, the need to effectively communicate the meaning of scores has become more prominent [6–9].

For some health domains, there are clear external standards with which to compare scores of a PRO measure. These comparisons aid interpretation. For example, when a PRO measure captures symptoms associated with a diagnostic disorder (e.g., major depressive disorder), clinician interview can serve as an external criterion. Some PRO measure scores (e.g., depressed mood) can be compared to clinical diagnoses using receiver operating characteristic curves and thresholds can be identified that optimize sensitivity and specificity.

Unfortunately, most PROs lack clear, "gold-standard" external criteria for comparison. In these cases thresholds may be defined statistically. However, statistical approaches can be limited with regard to their generalizability beyond the sample from which they were derived; further, statistical approaches do not confer meaning or clinical relevance unless paired with clinical interpretation or action. A reasonable approach is to apply multiple methods in gathering evidence to support judgments regarding recommended thresholds, a process called triangulation [10].

Our goal for this project was to develop clinically-relevant thresholds for three domains from the Patient-Reported Outcomes Measurement Information System® (PROMIS®): physical function, cognitive function, and sleep disturbance. These domains are important to people with cancer and therefore the clinicians who care for them and do not have established severity thresholds [11]. We refer to these as *provisional* thresholds because their clinical usefulness and generalizability will need to be assessed in specific contexts. We employed the "bookmarking" method to estimate score thresholds and apply qualitative labels to score ranges that can communicate severity levels. Bookmarking is a standard setting methodology used extensively in educational testing [12, 13]. This approach begins with a set of validated test questions ordered by level of difficulty. Using individual expert judgement and consensus-focused discussion, participants identify the location of an important threshold such as "proficiency" or passing. This is an exercise that applies meaningful labels (e.g., novice, intermediate, advanced, superior) to specific test scores. PRO measures quantify a symptom or function (e.g., depressive symptoms) that is best addressed in different ways depending upon its severity (e.g., psychoeducation, referral for treatment, psychiatric hospitalization). Consequently, bookmarking has been applied to measures of PROs to assign severity thresholds that may be tested, refined, and ultimately used to inform clinical action [14–17]. In this article, we describe an application of this methodology to identify PROMIS score thresholds in the context of cancer care and suggest next steps in the development of meaningful thresholds for use in clinical research and clinical practice.

Methods

Methods were modeled on previous PRO bookmarking studies [14–17] and are described in detail elsewhere [18]. The study was reviewed by the Northwestern University Institutional Review Board and determined to be exempt. All participants signed informed consent prior to participating.

Target measures

The target measures for this study were the PROMIS Item Bank v2.0—Physical Function [19, 20], Item Bank v2.0—Cognitive Function [21], and Item Bank v1.0—Sleep Disturbance [22]. These measures are scored by transforming raw scores to *T* scores that are normed on a sample that matched the U.S. general population with respect to age, sex, race/ethnicity, and education [23]. *T* scores have a mean of 50 and standard deviation of 10. High scores reflect more of the domain being assessed (i.e., high physical and cognitive function scores reflect better health; high sleep disturbance scores reflect poorer health).

PROMIS measures are calibrated using an item response theory (IRT) model [23, 24]. For each item, there is a single, most likely response associated with each possible PROMIS *T* score. For example, the response for the item, "My sleep was restless" is most likely to be "not at all" for individuals with low sleep disturbance and "very much" for individuals with high sleep disturbance. IRT-calibrated item banks such as the PROMIS measures are well suited to the bookmarking method because item banks have more items than traditional measures providing greater choice in selecting items for vignettes. Further, IRT calibration provides a model-based prediction of how individuals with different levels of PROs are most likely to respond to each item of the item bank. Thus each item can be associated with one response choice at every possible score level.

Procedures

Vignette construction

Vignettes were developed with five items. Collectively, they covered the range of possible scores for each item bank (physical function $T = 12.5–62.5$,

cognitive function $T = 22.5–57.5$, and sleep disturbance $T = 32.5–72.5$). Vignette locations were 1/2 standard deviation (SD) apart. Response probabilities for each item of the item banks were calculated using a custom program (available upon request) written using the R programming language v3.3.3 [25]. Vignettes are included in the supplementary file.

In selecting items for each vignette, we balanced a number of considerations: (1) maximized response variation within vignettes (e.g., all items within a vignette did not have the same response option), (2) no item repetition in adjacent vignettes, and (3) balanced subdomain content (e.g., within physical function include items about mobility, activities of daily living, and upper extremity function in a single vignette). A name was assigned to each vignette (e.g., Ms. Wright) using the most common surnames from the 2010 U.S. Census with gender alternating between adjacent vignettes (see Fig. 1).

Clinician participants

We identified oncology clinicians representing a range of professions (e.g., oncologist, advanced practice nurse, infusion nurse, clinical psychologist, social worker) and targeted inclusion of at least one person from each profession working in patient populations most likely to experience issues with physical function, cognitive function, and sleep disturbance. We also targeted clinicians with a range of tumor type expertise (e.g., gynecologic/breast cancers, prostate cancer, hematologic malignancies, etc.). The goal was to comprise a group that included males and females, a range of clinical disciplines, and a range of tumor specialties. Clinicians were contacted by email. Interested and eligible clinicians were given a description of the study and they provided verbal agreement to participate. Eligibility criteria included working as a clinical provider in oncology with at least 3 years of experience managing the selected functions and symptoms, treating at least 100 people with cancer, affiliation with the academic medical center, and available at the time of the focus group.

Ten female oncology clinicians agreed to serve as participants; one male clinician was recruited but dropped out prior to the study. Clinicians ranged in age from 31 to 47 years (mean = 35 years, 1 had missing data) and had 3–15 years of oncology experience (mean = 7.1, 1 missing data). The majority (80%) were White with 10% Asian-American and 10% African-American. Professions included oncology nurse ($n = 4$), mid-level provider ($n = 3$), oncologist ($n = 1$), clinical social worker ($n = 1$), and clinical psychologist ($n = 1$). Specialties included hematologic disorders, cancer survivorship, and gastrointestinal, genitourinary, breast, lung, and thoracic cancers.

Patient participants

Participating clinicians were asked for permission to approach their cancer patients with currently scheduled treatment or follow-up visits about participating in the patient-only focus group. The study coordinator reviewed the clinic schedule and approached eligible patients. Additionally, oncology social workers and clinical psychologists were provided with information about the study to share with potential participants. Interested potential participants were provided with a summary of the study and completed an eligibility screener. Eligibility criteria included having a cancer diagnosis (excluding basal cell and squamous cell skin cancer) in the previous 2 years, at least some difficulty with sleep disturbance, cognitive function, or physical function due to cancer or cancer treatment, receiving care at the cancer center, proficiency with written and spoken English, ability to send and receive email, and availability at the time of the focus group. Eligible patients signed an agreement and were provided with materials to complete before the focus group.

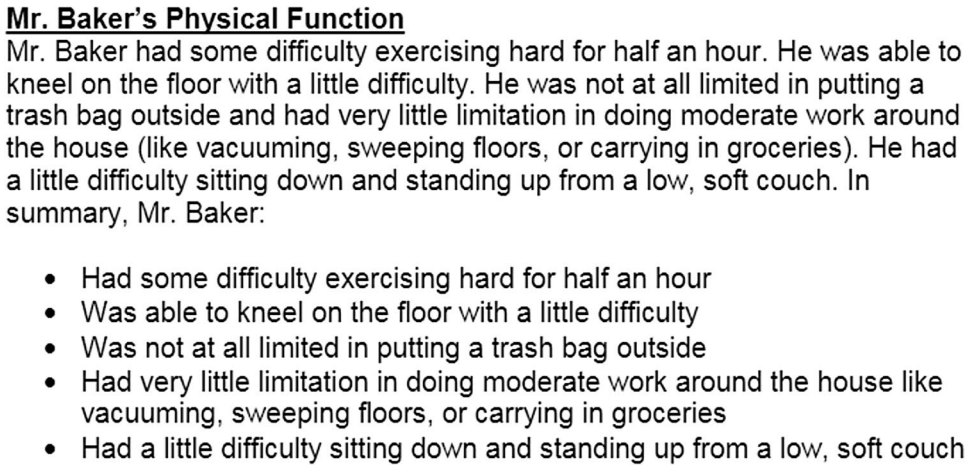

Fig. 1 Example physical function vignette

Eighteen people were approached to participate. Of these, twelve declined due to meeting logistics (e.g., working, live far away; $n=6$), not feeling well physically or emotionally ($n=3$), not interested ($n=2$), or the oncologist did not want the patient approached ($n=1$). Five women and one man participated in the focus group. Their ages ranged from 57 to 72 (mean = 62.7); 50% were African-American and 50% were White. None were currently partnered. Most (83%) completed a college or post-graduate degree. Two participants were retired and others were working part-time ($n=1$), working full-time ($n=1$), on disability ($n=1$), or unemployed ($n=1$). Cancer diagnoses included breast ($n=2$), lung, skin, angiosarcoma, and mantle cell lymphoma. Participants averaged about 20 months post-diagnosis (range 11 to 36 months). Half were currently receiving chemotherapy. Of those, one also was receiving hormonal treatment.

Study protocol

Clinician focus group The clinician focus group aims were to (a) identify terminology describing levels of severity, (b) validate vignettes of PROMIS items reflect different levels of severity, and (c) reach consensus on score thresholds distinguishing levels of severity. Clinicians were compensated $ 400 for their participation. After providing sociodemographic information, clinicians completed short forms for each of the three domains in order to orient them to the domains. A member of the research team presented an introduction to the rationale and methods used in bookmarking. Clinicians then engaged in a warm-up exercise in which they independently rank ordered desserts by level of fanciness [18].

For each of the three target domains, clinicians independently ranked vignettes according to their levels of severity. Rankings were recorded, reported to the group, and discussed until reaching consensus. The derived order was used as a check on the degree to which the vignettes communicated distinct and ordered levels of the target domain. Next, the moderator led a discussion of preferred terminology to describe patients' levels of the symptom or function. Discussion continued until consensus was reached.

Each clinician was given printed copies of the vignettes ordered from least to most severe based on their IRT-derived vignette locations. They also were given paper bookmarks labeled with the severity terms they had agreed upon earlier. Independently, clinicians placed a bookmark before the first vignette they believed no longer reflected the lowest severity term and continued with this procedure for the remaining bookmarks. They recorded and shared their bookmark locations. The moderator led the group in discussion of everyone's selected locations with the aim of reaching agreement; discussion continued until consensus was reached.

Patient focus group The patient focus group aims were to (a) validate vignettes of PROMIS items reflect different levels of severity, and (b) reach consensus on score thresholds distinguishing levels of severity. Patient participants were compensated $150 for their participation. The patient focus group was conducted after the clinician focus group. The procedures mirrored those of the clinician group with a few exceptions. Patient participants: (1) ordered the vignettes by severity prior to the group meeting and returned them via email, (2) completed a short form for each domain in advance of the group meeting, (3) were given the clinician-selected terms and asked to describe what the terms meant to them (e.g., "What makes a symptom 'severe'?"), and (4) at the end of bookmarking, provided feedback on the severity terms.

Data analysis

For each domain and each stakeholder group, we calculated the median vignette rank and compared this to the *T* score ranking. Final bookmarking results were used to define thresholds for each category, defined as the midpoint between the *T* score location of the two vignettes that shouldered the bookmark.

Results

Descriptive labels for score ranges

Clinicians

Clinicians struggled to identify useful labels for levels of symptom severity and dysfunction. Many terms were considered including "Problems," "Difficulty," "Issues," "Satisfaction," "Disturbance," "Manageability," and "Challenging" with a range of modifiers (within normal limits/mild/moderate/severe; none/little/a lot; no change needed/mild change needed/etc.). Some clinicians thought the terms should indicate whether a particular clinical action was needed. Others thought the labels should describe *levels of concern or problem* for the patient rather than *severity levels*. Some did not like this approach because patients might not be concerned about something their clinicians would be. The group expressed apprehension about possible negative reactions by patients to the term "severe." For the bookmarking exercise, clinicians agreed to utilize "within normal limits," "mild," "moderate," and "severe" to describe levels of all three domains.

Patients

Patients did not derive labels for use in bookmarking but were asked to define the labels supplied by the clinicians. They defined mild as "more of an irritation versus impairment" and "an annoyance." A moderate symptom was described as something that "hung around longer," "had greater suffering," was "harder to put out of your mind because you were aware of it at all times," was "intrusive," had more "discomfort," and was "not really bad, but not really good, but in the middle." Severe was described as having "a lot of fear around it," "a lot of symptoms," "unpleasant," "requiring immediate attention," "preventing doing things," and "intolerable."

After deriving consensus bookmark placements for all three domains using the provided severity labels, patients were asked to express their preferences for how ranges of symptoms and function should be labeled. Patients preferred terms that conveyed the treatability of a symptom or function (e.g., "severe but treatable"). Mild/moderate/severe were described as "non-action" words. One patient advocated for using numbers and not terms. Others thought the numbers would still require an interpretation guideline (e.g., 1–3 = mild) and that numbers alone are "scary" as they are similar to cancer staging. The term "impairment" generated both positive and negative reactions. Some believed the word "impairment" to be negative and stigmatizing and preferred not to be labeled as "impaired." Others thought the word "impairment" focused on abilities and, therefore, was less "scary" than other terms. The label "issues" was viewed favorably because it connoted the potential manageability of the symptom in contrast to something that "has been done to you". Similarly "manageable" was perceived as empowering and positive, although one patient felt it was not specific enough.

Independent ordering of vignettes by clinicians and patients

Table 1 reports the median, independent ordering of vignettes by individual patients and clinicians. Clinicians' median rankings of the cognitive function vignettes exactly matched the *T* score ranks. Their ranks for sleep disturbance and physical function were closely aligned with the exception of the reversal of two vignettes covering very poor outcomes. Overall, patients' rankings also aligned with the *T* score order, but there were more discrepancies as compared to those of clinicians. There also tended to be larger differences in patient ratings as compared to clinician ratings (e.g., patient cognitive function ratings in Table 1).

Independent bookmark placement compared to consensus placement

For sleep disturbance and cognitive function, no clinician's bookmarks completely matched the final group consensus; all clinicians moved at least one of their initial bookmarks in response to the group discussion. For physical function, four clinicians placed bookmarks identically to the group consensus.

Patients also were able to independently place bookmarks, although one participant did not use the "mild" category for cognitive function or physical function. No single patient's bookmark placement matched the final group consensus for cognitive function. For sleep disturbance, 2 patients matched the group and for physical function, 1 patient matched the group.

Table 1 Independent vignette order

T score	Sleep disturbance		*T* score	Cognitive function		Physical function	
	Clinician median order	Patient median order		Clinician median order	Patient median order	Clinician median order	Patient median order
32.5	1	1	62.5			1	1
37.5	2	3	57.5	1	1	2	2
42.5	3	2	52.5	2	2.5	3	3
47.5	4	4.5	47.5	3	2	4	4
52.5	5	5	42.5	4	4	5	6
57.5	6	6.5	37.5	5	5	6	5.5
62.5	7	7.5	32.5	6	6.5	7	6.5
67.5	9	8	27.5	7	8	9	9
72.5	8	8	22.5	8	6.5	8	8.5
			17.5			10	9.5
			12.5			11	11

Comparing clinician and patient bookmark placement

Clinicians' and patients' thresholds for sleep disturbance were identical (see Fig. 2). For cognitive and physical function, both groups set the same thresholds between "within normal limits" (WNL) and "mild". However, patients placed the other two thresholds higher than did clinicians. That is, compared to clinicians, patients judged lower levels of function less severely. This was particularly the case for the poorest levels of physical function. Whereas the clinicians' threshold defined physical function scores less than 30 as "severe", the patients' threshold defined severe as scores less than 20.

Discussion

In this study, standard setting bookmarking methods were applied to PROMIS measures for physical function, cognitive function, and sleep disturbance. This is the first study to evaluate patients' and clinicians' opinions regarding the semantic labels applied to ranges of symptoms and function. Though many options were discussed by both groups, there was no set of options that garnered unanimous support by either patients or clinicians. Both groups expressed preference for labels that referenced clinical implications. Some clinicians wanted labels that indicated whether a particular clinical action was needed; some patients preferred terms that conveyed the treatability of a symptom or function (e.g., "severe but treatable").

Participants were able to apply severity labels (within normal limits, mild, moderate, severe) to all domains and to reach consensus on threshold locations. For the different measures, few of patients' or clinicians' initial bookmarks matched the final group consensus. This suggests the consensus process was collaborative and not dominated by a single participant.

There was both consistency and variation between clinicians' and patients' consensus results. Sleep disturbance thresholds were identical; those for physical and cognitive function varied. When there was variance, patients consistently required more dysfunction than clinicians before assigning more severe labels. This result is consistent with findings from previous studies that compared patients' and clinicians bookmarking results. In one such study, individuals with Multiple Sclerosis and clinicians chose the same thresholds for Neuro-QOL measures of sleep disturbance and lower-extremity function. Compared to clinicians, patients required worse fatigue and lower-extremity function before classifying levels as more severe [15]. The same trend was observed in a bookmarking study in juvenile idiopathic arthritis (JIA) [17]. Interestingly, a bookmarking study in adults with rheumatic diseases found the opposite trend for physical function, sleep disturbance, and depression [16].

To our knowledge, this is the first study testing these measures in people with cancer. One study in rheumatic disease used similar methods with the same PROMIS measures [16]. For sleep disturbance, the rheumatic disease clinicians identified the same thresholds as both patients and clinicians in our study apart from using "severe" beginning at $T=60$ rather than $T=65$. For physical function, both

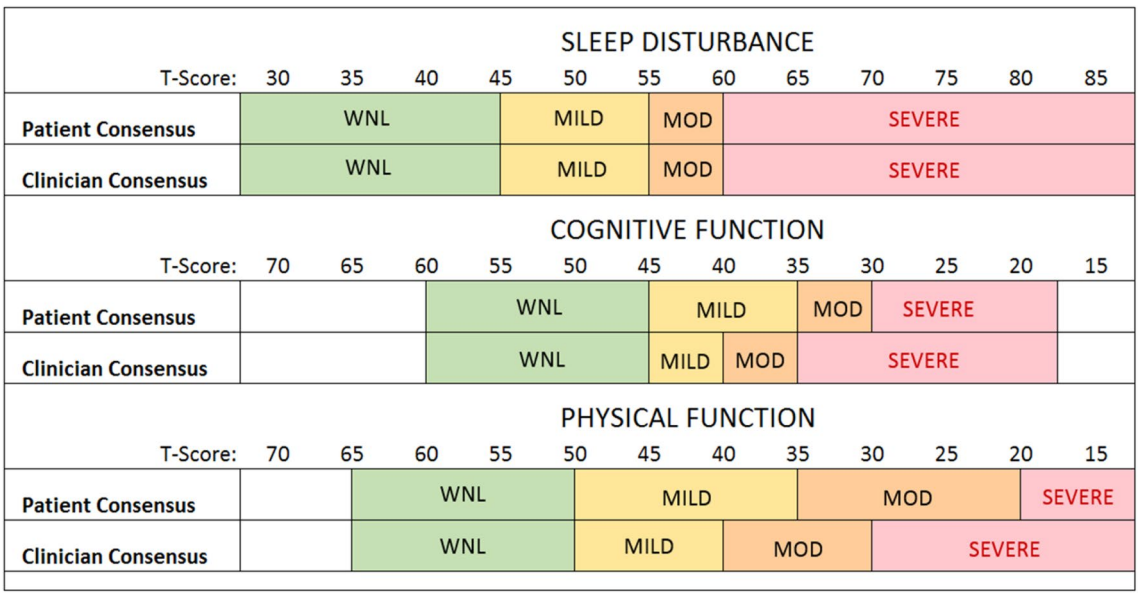

Fig. 2 Comparison of patients' and clinicians' thresholds for levels of symptoms and function

cancer patients and clinicians required more dysfunction than rheumatology patients and clinicians before assigning more severe labels. It is unclear to what extent these differences may be due to variability in small groups, differences across patient populations, use of the term "no problems" as the best possible label rather than "within normal limits," instruction to set thresholds for clinical action versus description alone, patient adaptation, or other factors.

This study has several limitations. First, our sample size was small. Although typical for qualitative research, small sample sizes do not represent the range of cancer care providers and patients. Future studies should include additional physicians (e.g., oncologists, palliative care) and male participants. Second, a general limitation when identifying thresholds among both clinicians and patients is that their decisions inherently rely upon different contextual information. The frame of reference for clinicians making judgments is likely to be their patient population, opinions of trusted colleagues, and the published literature. The patient's frame of reference is more personal, including their own experiences and those of other patients they may know. Both of these perspectives are informative, although they use different information. Methods need to be devised that coherently balance these different perspectives and aid interpretation. Other limitations pertain to small group consensus processes. Groups may be swayed by the opinions of more vocal members or be overly influenced by the expressed goal of reaching consensus. The extent to which these limitations influenced our study results will become clearer with future studies that replicate bookmarking methods for the target measures.

Conclusions

This study produced clinical thresholds for action for PROMIS *T* scores of physical function, cognitive function, and sleep. Though patients had more variability with the *T* score ranking than clinicians, both patients and clinicians were able to perceive vignettes as representing different levels—evidenced by their ability to rank order vignettes by severity. These findings support the use of bookmarking methods in evaluating PRO score interpretation.

The generalizability and reliability of bookmarking results needs to be evaluated in future studies across a range of diagnoses, treatments, and patient factors. Future research can also help evaluate clinician and patient preferences for semantic labels for interpreting levels of a PRO. Although the labels used in this study have interpretative value, it is possible that setting thresholds to differentiate diagnostic severity may produce different values than setting thresholds for clinical action. The results of this study provide a useful starting point for these and other initiatives.

Acknowledgements This study was funded by the National Institutes of Health Grant U2C CA186878.

References

1. Chen, J., Ou, L., & Hollis, S. J. (2013). A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. *BMC Health Services Research, 13,* 211–211.
2. Howell, D., Molloy, S., Wilkinson, K., Green, E., Orchard, K., Wang, K., et al. (2015). Patient-reported outcomes in routine cancer clinical practice: A scoping review of use, impact on health outcomes, and implementation factors. *Annals of Oncology, 26*(9), 1846–1858.
3. Wagner, L. I., Schink, J., Bass, M., Patel, S., Diaz, M. V., Rothrock, N., et al. (2015). Bringing PROMIS to practice: Brief and precise symptom screening in ambulatory cancer care. *Cancer, 121*(6), 927–934.
4. Seneviratne, M. G., Bozkurt, S., Patel, M. I., Seto, T., Brooks, J. D., Blayney, D. W., et al. (2019). Distribution of global health measures from routinely collected PROMIS surveys in patients with breast cancer or prostate cancer. *Cancer, 125*(6), 943–951.
5. Lohr, K. N., Aaronson, N. K., Alonso, J., Audrey Burnam, M., Patrick, D. L., Perrin, E. B., et al. (1996). Evaluating quality-of-life and health status instruments: Development of scientific review criteria. *Clinical Therapeutics, 18*(5), 979–992.
6. Snyder, C. F., Smith, K. C., Bantug, E. T., Tolbert, E. E., Blackford, A. L., Brundage, M. D., et al. (2017). What do these scores mean? Presenting patient-reported outcomes data to patients and clinicians to improve interpretability. *Cancer, 123*(10), 1848–1859.
7. Cappelleri, J. C., & Bushmakin, A. G. (2014). Interpretation of patient-reported outcomes. *Statistical Methods in Medical Research, 23*(5), 460–483.
8. Given, B., Given, C. W., Sikorskii, A., Jeon, S., McCorkle, R., Champion, V., et al. (2008). Establishing mild, moderate, and severe scores for cancer-related symptoms: How consistent and clinically meaningful are interference-based severity cut-points? *Journal of Pain and Symptom Management, 35*(2), 126–135.
9. Palos, G. R., Mendoza, T. R., Mobley, G. M., Cantor, S. B., & Cleeland, C. S. (2006). Asking the community about cutpoints used to describe mild, moderate, and severe pain. *The Journal of Pain, 7*(1), 49–56.
10. Coon, C. D., & Cook, K. F. (2018). Moving from significance to real-world meaning: Methods for interpreting change in clinical outcome assessment scores. *Quality of Life Research, 27*(1), 33–40.
11. Basch, E., Reeve, B. B., Mitchell, S. A., Clauser, S. B., Minasian, L. M., Dueck, A. C., et al. (2014). Development of the National Cancer Institute's Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). *JNCI: Journal of the National Cancer Institute, 106*(9), 244.
12. Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93–106.
13. Perie, M. (2005). Angoff and Bookmark methods. Workshop presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.
14. Cella, D., Choi, S., Garcia, S., Cook, K. F., Rosenbloom, S., Lai, J.-S., et al. (2014). Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. *Quality of Life Research, 23*(10), 2651–2661.
15. Cook, K. F., Victorson, D. E., Cella, D., Schalet, B. D., & Miller, D. (2015). Creating meaningful cut-scores for Neuro-QOL

measures of fatigue, physical functioning, and sleep disturbance using standard setting with patients and providers. *Quality of Life Research, 24*(3), 575–589.
16. Nagaraja, V., Mara, C., Khanna, P. P., Namas, R., Young, A., Fox, D. A., et al. (2018). Establishing clinical severity for PROMIS® measures in adult patients with rheumatic diseases. *Quality of Life Research, 27*(3), 755–764.
17. Morgan, E. M., Mara, C. A., Huang, B., Barnett, K., Carle, A. C., Farrell, J. E., et al. (2017). Establishing clinical meaning and defining important differences for Patient-Reported Outcomes Measurement Information System (PROMIS®) measures in juvenile idiopathic arthritis using standard setting with patients, parents, and providers. *Quality of Life Research, 26*(3), 565–586.
18. Cook, K., Cella, D., & Reeve, B. (2019). PRO-bookmarking to estimate clinical thresholds for patient-reported symptoms and function. *Medical Care, 57,* S13–S17.
19. Rose, M., Bjorner, J. B., Becker, J., Fries, J., & Ware, J. (2008). Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of Clinical Epidemiology, 61*(1), 17–33.
20. Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. E. (2014). The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology, 67*(5), 516–526.
21. Lai, J. S., Wagner, L. I., Jacobsen, P. B., & Cella, D. (2014). Self-reported cognitive concerns and abilities: two sides of one coin? *Psycho-Oncology, 23*(10), 1133–1141.
22. Buysse, D. J., Yu, L., Moul, D. E., Germain, A., Stover, A., Dodds, N. E., et al. (2010). Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. *Sleep, 33*(6), 781–792.
23. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63*(11), 1179–1194.
24. Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45*(5), S22–S31.
25. R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.