PROMIS 4-Item Measures and Numeric Rating Scales Efficiently Assess SPADE Symptoms Compared to Legacy Measures

K Kroenke, MD;[1,2] TE Stump, MS;[3] J Kean, PhD;[4] TL Talib, PhD;[2]
DA Haggstrom, MD;[1,2,5] PO Monahan, PhD[3]

[1]Division of General Internal Medicine and Geriatrics, Indiana University School of Medicine, Indianapolis, IN, USA

[2]Regenstrief Institute, Inc., Indianapolis, IN, USA

[3]Department of Biostatistics, Indiana University Fairbanks School of Public Health and School of Medicine, Indianapolis, IN, USA

[4]Department of Population Health Sciences, University of Utah School of Medicine, Salt Lake City, Utah, USA

[5]VA Center for Health Information and Communication, Indianapolis, IN, USA

Corresponding Author: Kurt Kroenke, MD, Regenstrief Institute, Inc., 1101 West 10th St., Indianapolis, IN 46202. Phone: 317-274-9046 Fax: 317-274-9304. E-mail: kkroenke@regenstrief.org

Funded by Patient-Centered Outcomes Research Institute (PCORI) ME-1403-12043. The sponsor had no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Declarations of interest: none

Abstract

Objective: The 5 SPADE symptoms (**s**leep, **p**ain, **a**nxiety, **d**epression, and low **e**nergy/fatigue) are among the most prevalent and disabling symptoms in clinical practice. This study evaluates the minimally important difference (MID) of PROMIS measures and their correspondence with other brief measures to assess SPADE symptoms.

Study Design and Setting: 300 primary care patients completed a 4-item PROMIS scale, a numeric rating scale (NRS), and a non-PROMIS legacy scale for each of the 5 SPADE symptoms. Optimal NRS cutpoints were examined and cross-walk units for converting legacy measure scores to PROMIS scores were determined. PROMIS scores corresponding to standard deviation (SD) and standard error of measurement (SEM) changes in legacy scores were used to estimate MID.

Results: At an NRS ≥ 5, the mean PROMIS T-score exceeded 55 (the operational threshold for a clinically meaningful symptom) for each SPADE symptom. Correlations were high (0.70-0.86) between each PROMIS scale and its corresponding non-PROMIS legacy scale. Changes in non-PROMIS legacy scale scores of 0.35 SD and 1 SEM corresponded to mean PROMIS T-scores of 2.92 and 3.05 across the 5 SPADE symptoms, with changes in 0.2 and 0.5 SD corresponding to mean PROMIS T-scores of 1.67 and 4.16

Conclusion: A 2-step screening process for SPADE symptoms might use single-item NRS scores, proceeding to PROMIS scales for NRS scores ≥ 5. A PROMIS T-score change of 3 points represents a reasonable MID estimate, with 2 to 4 points approximating lower and upper bounds.

Key Words: psychometrics; pain; depression; anxiety; sleep; fatigue

Running Title: Measuring Symptoms with PROMIS Scales

Word Count: 3254

Trial Registration Number: ClinicalTrials.gov ID: NCT02383862

What is new?

Key findings

- Both single-item numeric rating scales (NRS) and 4-item PROMIS scales can efficiently assess the 5 SPADE symptoms (i.e., sleep problems, pain, anxiety, depression, and low energy/fatigue)
- A 3-point change in the PROMIS T-score represents a minimally important difference (MID), with 2 to 4 points representing the lower and upper MID bounds.

What this adds to what was known?

- An NRS score ≥ 5 is a useful screening cut-point for SPADE symptoms.

What is the implication and what should change now?

- Ultra-brief public domain scales can be used to detect and monitor common symptoms which are highly prevalent and burdensome in clinical practice.

1. Introduction

The SPADE pentad – ***s***leep problems, ***p***ain, ***a***nxiety, ***d***epression, low ***e***nergy/fatigue – comprises a frequently undertreated group of symptoms that is prevalent across most medical and mental disorders [1, 2]. Moreover, these 5 symptoms are often chronic, co-occur with one another, and cause substantial impairment. The ubiquitous nature and clinical significance of these symptoms is further substantiated by the fact that the SPADE pentad constitutes 5 of the 7 domains included in the PROMIS (Patient-Reported Outcomes Measurement Information System) 29-item, 43-item, and 57-item profiles (the other 2 domains are physical function and satisfaction with social roles). PROMIS scales are a widely used set of patient-reported outcome measures developed with NIH support [3, 4], and the PROMIS profiles consist of 4-item, 6-item, or 8-item scales to assess 7 domains that are cross-cutting across a wide range of diseases and health conditions. Besides the multi-domain profiles, there are larger item banks from which both the profiles as well as domain-specific short forms are derived.[3] Another unique aspect of PROMIS scales is that unlike many legacy scales which are developed using classical test theory, PROMIS scales were developed using item response theory, the potential advantages of which are described elsewhere.[5]

Assessing patient-reported outcomes (PROs) is becoming an increasing priority in both clinical practice and research [6-8]. Measurement-based care is an essential component of both detecting as well as improving outcomes for depression, anxiety, pain and other symptom-based conditions [9, 10]. Indeed, embedding PROs in electronic health care records is gaining momentum and intended to foster enhanced symptom management [11-15]. Similarly, assessing symptoms as primary or secondary outcomes in clinical research is facilitated by brief, validated, self-reported measures [7, 16].

In this study, we compare the PROMIS 4-item scales for the SPADE symptoms to single-item numeric rating scales (NRS) as well as ultra-brief (2-4 items) legacy non-PROMIS scales. Specifically, our aims were to determine: 1) the optimal NRS cutpoint in screening for clinically relevant SPADE symptoms; 2) cross-walk units for converting non-legacy scale scores to PROMIS scores; 3) preliminary estimates of the minimally important difference (MID) in PROMIS scores for the SPADE symptoms.

2. Methods

2.1. Study procedures

In this prospective, 2-arm randomized clinical trial, a research assistant recruited patients over a 13-month period from primary care (internal medicine and family medicine) clinics located within an urban academic setting. Upon checking in for their clinic visit, patients were asked to complete a 5-item symptom screener adapted from the MD Anderson Symptom Inventory [17] consisting of a numeric rating scale (NRS) for the severity over the past few weeks of each SPADE symptom that ranged from 0 (“not present”) to 10 (“as bad as you can imagine”). Patients who had at least 1 SPADE symptom with a NRS severity score ≥ 4 and who provided informed consent completed the 20-item PROMIS questionnaire (4-items scales for each of the SPADE symptoms) on an iPad tablet. Study participants were then randomized to a feedback group, in which their clinician received a graph of their symptom scores, or a control group. Three months after the baseline visit, the PROMIS scales and secondary outcome measures were administered using either paper or web-based surveys (patient preference). A 3-month follow-up assessment was chosen for the main trial in order to allow sufficient time to determine if feedback had improved symptom outcomes as a result of clinical actions and/or natural history. Non-respondents were contacted by phone to complete the survey by interview. The study was approved by Indiana University’s institutional review board. Further details of the SPADE trial have been previously described [2].

2.1. Measures

The PROMIS profile-29 includes 4-item scales for 7 domains; 5 of these domains were used for this study – sleep, pain, anxiety, depression, and fatigue – as they reflected the SPADE symptoms [3, 4]. PROMIS scales provide 5 response options for each item which allow patients to indicate the severity of symptoms in the past 7 days. Each PROMIS scale provides a raw score, ranging from 4 to 20. Raw scores are converted to T-scores using the PROMIS conversion tables. T-scores allow for standardized comparison across symptoms and calculation of a composite symptom score. A T-score of 50 on each PROMIS symptom scale represents the general population norm, and each 10-point deviation represents one standard deviation (SD) from the population norm. For the purposes of the study, a cutpoint of ≥ 55 was used to represent a clinically-elevated symptom score as this is ≥ 0.5 SD worse than the population norm, which is traditionally considered a moderate effect size [18]. Also, this cutpoint was used in a previous study examining the SPADE pentad [1].

Non-PROMIS legacy measures for each SPADE symptom were administered at 3-month follow-up to compare PROMIS scales to established, well-validated measures. Ultra-brief measures (operationally defined as 1 to 4 items [19, 20]) were used to parallel the brevity of the 4-item PROMIS measures as well as to minimize respondent burden.

The *Pittsburgh Insomnia Rating Scale* (PIRS) was designed to evaluate the severity of insomnia in clinical trials. The abbreviated 2-item PIRS assesses sleep disturbance and satisfaction with sleep in the past week on a 4-point (0 to 3) scale. Possible scores on the PIRS-2 range from 0 to 6, with higher scores reflecting more severe sleep problems. PIRS has demonstrated good internal consistency, test-retest reliability and concurrent validity with standard measures of insomnia, such as the Insomnia Severity Index and Pittsburgh Sleep Quality Index [21, 22].

The *PEG* scale (**P**ain intensity, **E**njoyment of life, and **G**eneral activity) is a 3-item version of the Brief Pain Inventory (BPI) that assesses pain in the past week and demonstrates reliability, validity and sensitivity to change comparable to the 11-item BPI as well as other standard pain measures [23-25]. PEG scores range from 0 to 10, with higher scores indicating more severe pain intensity and interference.

The PHQ-4 is a well-validated 4-item scale that consists of the *GAD-2* anxiety scale and the *PHQ-2* depression scale. Using a 4-point Likert scale (0 = not at all; 3 = nearly every day), patients report how often they have experienced anxiety and depressive symptoms over the past 2 weeks. Both the GAD-2 and PHQ-2 scores range from 0 to 6, with higher scores indicating more severe anxiety and depression, respectively. Both the PHQ-4 as a composite mood scale and the GAD-2 and PHQ-2 as anxiety and depression subscales have shown strong internal consistency, construct and criterion validity, and sensitivity to change [26-31].

The Short-Form-36 Health Survey (*SF-36) vitality* scale assesses the frequency with which patients have experienced symptoms of fatigue or low energy over the past week on a 5 point Likert scale (1 = all of the time; 5 = none of the time). Scores are transformed to a 0 to 100-point scale, with lower scores representing greater fatigue. The SF-36 vitality scale is one of the best-validated brief measures of fatigue [32, 33].

2.3. Statistical analysis

Means and standard deviations were determined for all scale scores. Correlations between each PROMIS symptom score and its corresponding NRS and non-PROMIS legacy measure score were calculated. The relationship between NRS and baseline PROMIS scores were graphically examined. Also, the operating characteristics of varying NRS cutpoints for clinical SPADE symptoms (i.e., PROMIS T-score ≥ 55) were calculated, including sensitivity, specificity, positive and negative predictive values, and Youden's index ([sensitivity + specificity] – 1). Additionally, the area under the curve (AUC) for NRS was examined using receiver operating curve analyses. Finally, cross-walking (calibrating) of scale scores were conducted by using linear regression to determine how much the PROMIS T-score changed for a 0.2, 0.35, and 0.5 standard deviation (SD) change in the non-PROMIS legacy measure as well as a 1 standard error of measurement (SEM) change; the latter is calculated as SD times the square root of (1 – Cronbach's alpha). Score changes of 0.35 and 1 SEM are considered two valid point estimates of the minimally important difference (MID) for a scale, with 0.2 and 0.5 SD representing lower and upper MID bounds.[34] An MID is defined as the smallest difference in a scale score that patients perceive as important, either beneficial or harmful, and that would prompt a clinician to consider a change in the patient's management.

3. Results

3.1. Patient sample

Of 419 patients screened in clinic, 374 (89%) screened positive for at least 1 of the 5 SPADE symptoms. Thirty eligible patients declined. Another 44 eligible patients were interested and willing to participate, but due to inadequate time were unable to complete the enrollment process prior to seeing their doctor. Thus, 300 patients completed the 20-item PROMIS measures. Symptom screening scores did not significantly differ between patients who declined, those who enrolled, and those who were interested but unable to complete enrollment.

Follow-up data was collected 3 months after the baseline visit from 256 (85.3%) of the study participants. Compared to participants with follow-up data, the 44 participants without follow up data were younger (41.6 years vs. 50.7 years, $P < 0.001$). Otherwise, they were similar with regard to recruitment site, sex, race, education and baseline PROMIS composite T-score, indicating that follow-up data were largely missing at random.

Participants had a mean (SD) age of 49.4 (14.4) years; 215 (71.7%) were women; and race was white in 135 (45.0%), black in 148 (49.3%), and other in 17 (5.7%). Education was high school or less in 136 (53.3%), some college or trade school in 85 (33.3%), and a college degree or greater in 34 (13.3%).

3.2. Scale scores

Table 1 summarizes the NRS and baseline PROMIS scores (n =300), and follow-up PROMIS and legacy scale scores (n = 256). Mean baseline PROMIS scores indicate moderate symptom severity across the 5 SPADE symptoms. Participants largely had multiple SPADE symptoms; the proportion with 0, 1, 2, 3, 4 and 5 clinically significant symptoms (T-score ≥55) at baseline was 5%, 11%, 13%, 18%, 21%, and 31%, respectively. There were no significant differences in baseline characteristics between the feedback and control group. Also, changes in PROMIS scores at 3 months were similar in the feedback and control groups.[2]

3.3. Association between NRS and PROMIS scores

NRS and baseline PROMIS T-scores were strongly correlated, being 0.78 for sleep, 0.73 for pain and depression, 0.71 for fatigue, and 0.67 for anxiety. The **Figure** displays the strong linear association between NRS and PROMIS scores. Notably, at a NRS cutpoint of ≥ 5, the mean PROMIS T-score reached a clinical threshold of 55 or greater for all 5 SPADE symptoms.

The operating characteristics for NRS cutpoints of 4, 5, and 6 are summarized in **Table 2**. Although there was some variability across symptoms, sensitivity and negative predictive value tended to be higher at a cutpoint ≥ 5, whereas specificity and positive predictive value tended to be higher at a cutpoint ≥ 6. Compared to a cutpoint ≥ 6, a cutpoint ≥ 5 produced a higher Youden's index for 2 symptoms and a lower Youden's index for 3 symptoms.

3.4. Correlation and Cross-walking of PROMIS with Non-PROMIS Legacy Measures

Pearson correlations between PROMIS T-scores and non-PROMIS legacy measure scores at 3 months are summarized in **Table 3**. As expected, each PROMIS symptom scale correlates most strongly with the same-symptom non-PROMIS legacy scale (e.g., PROMIS pain scale with PEG pain scale); these same-symptom scale correlations ranged from .70 to .86. The next highest correlations were between depression and anxiety scales (.67 to .68).

Table 4 shows PROMIS scores cross-walked to non-PROMIS legacy scale scores. The mean MID point estimates for PROMIS T-scores across the 5 symptoms using 0.35 SD and 1 SEM change in the legacy scale scores were 2.92 and 3.05, respectively. The mean PROMIS T-score changes for a 0.2 and a 0.5 SD change in the legacy scales were 1.67 and 4.16. Thus, 3 points seems a reasonable MID estimate for PROMIS T-scores when assessing the SPADE symptoms, with 2 to 4 points representing the lower and upper bounds.

4. Discussion

SPADE symptoms were highly prevalent in our primary care sample and frequently co-occurred with one another. A single-item 0 to 10 numeric rating score was strongly associated with PROMIS scores, and an NRS cutpoint ≥ 5 proved to be a reasonable screening threshold for potential clinically meaningful symptoms (i.e., PROMIS T-score ≥ 55). Using non-PROMIS legacy scales, our data suggest that a PROMIS T-score change of 3 may represent a reasonable point estimate of a minimally important difference (MID) with 2 and 4 approximating the lower and upper bounds.

Evaluation of operating characteristics in Table 2 suggest that an NRS cutpoint of either 5 or 6 might be appropriate, depending upon the desired balance between sensitivity and specificity as well as the purpose of assessment (e.g., universal screening, suspicious clinical findings, treatment monitoring). Graphical examination (Figure 1) supports an NRS cutpoint of 5 in that the mean PROMIS T-scores for all 5 SPADE symptoms are 55 or higher at this cutpoint. Previous studies also support a threshold of 5 on a 0 to 10 NRS, or 50 on a 0 to 100 NRS [35, 36].

Our NRS findings have a couple practical implications. First, in some clinical settings, a single-item per SPADE symptom measure could be administered followed by the completion of PROMIS or other multi-item scales for the subset of individuals who screen positive on the NRS. Second, in research studies where one or more of the SPADE symptoms are secondary or tertiary (rather than primary) outcomes, an NRS might serve as a surrogate symptom marker to reduce respondent burden. The validity and utility of single-item symptom and quality of life scales has been substantiated by others [17, 35-38]. The tradeoffs in using single-item vs. multi-item scales has also been discussed [20, 39].

The MID of several PROMIS symptom scales has ranged from 2 to 5 points in patients with chronic pain [34, 40-42] or cancer [43]. In three of these studies, multiple methods were used to estimate MID [34, 42, 43]. Similarly, we found a 2 to 4-point MID range for the SPADE symptoms in a primary care sample. Importantly, the MID was similar regardless of the particular SPADE symptom or the fact that a different legacy measure was used as the MID anchor for each of the 5 symptoms. This finding, coupled with prior research, enhances confidence in a 2 to 4-point MID estimate for the PROMIS T-score in assessing SPADE symptoms.

Two operational decisions regarding PROMIS measures are: Should the scales be administered using computer-adaptive testing (CAT) or fixed forms, and should clinicians be provided raw or T-scores? Whereas CAT draws from a much larger data bank of items, it usually allows reliable scores to be achieved from 8 or fewer items by tailoring subsequent item selection to a particular patient’s previous responses. The benefits of CAT may vary with the size of the item bank as well as the trait range in the target population (e.g., depression may be more severe and skewed in mental health specialty settings than in primary care). The fixed form scales typically range from 4 to 8 items that do not vary but were provided by PROMIS developers for the many settings where CAT administration is not yet feasible. Since fixed forms provide reliable scores that may not differ substantially from CAT-derived scores, at least for clinical decision-making, the decision of whether to use fixed-forms vs. CAT can be tailored to the health care system’s setting and resources [3, 44-48], including the degree to which patient portals enabling self–administration of such measures are implemented among its population. Regarding raw vs. T-scores, the latter are preferable since for all scales, a T-score of 50 is the population norm and every 10 points is one standard deviation. However, conversion of raw to T-scores requires the use of tables which if done manually requires extra clinician time. In our study, this task was achieved by electronic administration of the PROMIS scales with programming that converted raw to T-scores.

There are also some practical questions regarding the assessment of patient-reported outcomes (PROs) in clinical practice. First, should single or multiple symptoms be targeted? Expanding the number of symptoms routinely assessed increases the amount of information provided to clinicians who already have many competing demands during time-limited primary care encounters. Second, should all patients be screened or only those at higher risk for problematic symptoms? Although universal screening can detect symptoms not volunteered by patients, it might also identify symptoms not warranting treatment, which could lead to extended office visit time or unnecessary testing. Third, how frequently should patients be assessed? Options include PRO administration for new patients, annual screening, assessment at every visit, or some other interval based upon initial screening scores, patient risk factors or comorbidity, or patient preferences.

Fourth, should assessment be aimed at symptom detection or treatment monitoring? Detection focuses on cross-sectional screening whereas monitoring requires longitudinal re-assessment in patients with initially high scores to determine if improvement has occurred and, if not, whether adjustments in treatment are warranted. Fifth, should PRO administration occur in-clinic or could it be completed at home via the Internet or telephone? Both methods have been automated in previous studies with high completion rates and patient satisfaction [49, 50].

Our study has several limitations. First, the trial was conducted in academic urban primary care clinics providing care to an underserved population that also includes a larger proportion of female and African-American patients relative to the U.S. population. Because PROMIS scores are population normed, this might have some influence on our findings and justifies further study in samples with differing demographics. Second, we tested the shortest PROMIS fixed scale (4 items per SPADE symptom rather than PROMIS scales ranging from 6 to 8 items). However, research has shown high correlations, similar MIDs, and comparable responsiveness of the fixed scales regardless of length [34, 45, 51]. Third, we used two anchor-based cross-sectional methods for estimating MID (i.e., SD and SEM changes on legacy scales) in a single primary care sample rather than triangulating multiple cross-sectional and longitudinal methods across a variety of populations, health conditions, and trait ranges [34, 43, 52, 53]. Fourth, a T-score ≥ 55 was used as a potential clinically relevant threshold for all 5 SPADE symptoms; however, the limited research on clinical thresholds for PROMIS scores suggest that thresholds might vary somewhat based upon the specific symptom, clinical disorder, and population sample.[54]

5. Conclusions

Capturing pre-visit SPADE scores is feasible with single-item or ultra-brief symptom measures. Because feedback alone, however, has had a weak effect on symptom outcomes in both our trial and previous studies[2, 55], potential priorities for future research might include:

1) Testing a stepped-care approach [56] with symptom self-management (ideally web-based) as step 1, nurse care management (much of which could be telephonic) augmented by automated symptom monitoring as step 2, and primary care physician engagement as step 3.

2) Eliciting patient preferences to decide if symptom treatment is desired [57, 58] and, if so, how multiple symptoms might be prioritized for treatment.

3) Developing efficient symptom-focused bedside evaluations for common symptoms, as have been developed for dizziness [59] and low back pain [60].

4) Implementing communication strategies (including brief scripts) for addressing symptom-specific concerns and avoiding unnecessary testing [61].

5) Studying the best way to treat multiple symptoms including treatments that may be effective across multiple types of symptoms [61] or sequential treatment focusing on the most severe symptom, most important to the patient, or most likely to respond to treatment.

Efficient PRO assessment coupled with effective management strategies is a desirable approach to optimizing outcomes in patients with SPADE and other bothersome symptoms.

References

[1] Davis LL, Kroenke K, Monahan P, Kean J, Stump TE. The SPADE Symptom Cluster in Primary Care Patients With Chronic Pain. Clin J Pain. 2016;32:388-93.

[2] Kroenke K, Talib TL, Stump TE, Kean J, Haggstrom DA, DeChant P, et al. Incorporating PROMIS symptom measures into primary care practice-a randomized clinical trial. J Gen Intern Med. 2018;33:1245-52.

[3] Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol. 2010;63:1179-94.

[4] Fischer F, Gibbons C, Coste J, Valderas JM, Rose M, Leplege A. Measurement invariance and general population reference values of the PROMIS Profile 29 in the UK, France, and Germany. Qual Life Res. 2018;27:999-1014.

[5] Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38:II28-42.

[6] Black N, Burke L, Forrest CB, Sieberer UH, Ahmed S, Valderas JM, et al. Patient-reported outcomes: pathways to better health, better services, and better societies. Qual Life Res. 2016;25:1103-12.

[7] Snyder CF, Jensen RE, Segal JB, Wu AW. Patient-reported outcomes (PROs): putting the patient perspective in patient-centered outcomes research. Med Care. 2013;51:S73-9.

[8] Basch E. Patient-reported outcomes - harnessing patients' voices to improve clinical care. N Engl J Med. 2017;376:105-8.

[9] Kroenke K, Unutzer J. Closing the false divide: sustainable approaches to Integrating mental health services into primary care. J Gen Intern Med. 2017;32:404-10.

[10] Kroenke K. Pain measurement in research and practice. J Gen Intern Med. 2018;33:7-8.

[11] Glasgow RE, Kaplan RM, Ockene JK, Fisher EB, Emmons KM. Patient-reported measures of psychosocial issues and health behavior should be added to electronic health records. Health Affairs. 2012;31:497-504.

[12] Snyder CF, Aaronson NK, Chouchair AK, Elliott TE, Greenhalgh J, Halyard MY, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. Qual Life Res. 2012;21:1305-14.

[13] Greenhalgh J. The applications of PROs in clinical practice: what are they, do they work, and why? Qual Life Res. 2009;18:115-23.

[14] Kotronoulas G, Kearney N, Maguire R, Harrow A, Di Domenico D, Croy S, et al. What is the value of the routine use of patient-reported outcome measures toward improvement of patient outcomes, processes of care, and health service outcomes in cancer care? A systematic review of controlled trials. J Clin Oncol. 2014;32:1480-501.

[15] Kroenke K, Monahan PO, Kean J. Pragmatic characteristics of patient-reported outcome measures are important for use in clinical practice. J Clin Epidemiol. 2015;68:1085-92.

[16] Wu AW, Kharrazi H, Boulware L, Snyder CF. Measure once, cut twice-adding patient-reported outcome measures to the electronic health record for comparative effectiveness research. J Clin Epidemiol. 2013;8:S12-S20.

[17] Cleeland CS, Mendoza TR, Wang XS, Chou C, Harle MT, Morrissey M, et al. Assessing symptom distress in cancer patients: the M.D. Anderson Symptom Inventory. Cancer. 2000;89:1634-46.

[18] Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care. 1989;27:S178-S89.

[19] Mitchell AJ, Coyne JC. Do ultra-short screening instruments accurately detect depression in primary care? A pooled analysis and meta-analysis of 22 studies. Br J Gen Pract. 2007;57:144-51.

[20] Hays RD, Reise S, Calderon JL. How much is lost in using single items? J Gen Intern Med. 2012;27:1402-3.

[21] Moul D, Pilkonis P, Miewald J, Carey T, Buysse D. Preliminary study of the test-retest reliability and concurrent validities of the Pittsburgh Insomnia Rating Scale (PIRS). Sleep: AMER ACAD SLEEP MEDICINE 6301 BANDEL RD, STE 101, ROCHESTER, MN 55901 USA; 2002. p. A246-A7.

[22] Veqar Z, Moiz JA, Hussain ME. Psychometric Analysis of the Pittsburgh Insomnia Rating Scale among University Population of Poor Sleepers in India. N Am J Med Sci. 2014;6:161-7.

[23] Krebs EE, Bair MJ, Damush TM, Tu W, Wu J, Kroenke K. Comparative responsiveness of pain outcome measures among primary care patients with musculoskeletal pain. Med Care. 2010;48:1007-14.

[24] Kroenke K, Theobald D, Wu J, Tu W, Krebs EE. Comparative responsiveness of pain measures in cancer patients. J Pain. 2012;13:764-72.

[25] Krebs EE, Bair MJ, Damush TM, Wu J, Sutherland J, Kroenke K. Development and initial validation of a 3-item Brief Pain Inventory. JGIM. 2008;23:278-9.

[26] Kroenke K, Spitzer RL, Williams JBW, Lowe B. An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics. 2009;50:613-21.

[27] Kroenke K, Spitzer RL, Williams JB, Lowe B. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. Gen Hosp Psychiatry. 2010;32:345-59.

[28] Lowe B, Wahl I, Rose M, Spitzer C, Glaesmer H, Wingenfeld K, et al. A 4-item measure of depression and anxiety: Validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. J Affect Disord. 2009.

[29] Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care. 2003;41:1284-92.

[30] Lowe B, Kroenke K, Grafe K. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). J Psychosom Res. 2005;58:163-71.

[31] Kroenke K, Spitzer RL, Williams JBW, Monahan PO, Lowe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med. 2007;146:317-25.

[32] McHorney CA, Ware JE, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care. 1993;31:247-63.

[33] O'Connor PJ. Evaluation of four highly cited energy and fatigue mood measures. J Psychosom Res. 2004;57:435-41.

[34] Chen CX, Kroenke K, Stump T, Kean J, Carpenter JS, Krebs EE, et al. Estimating minimally important differences for the PROMIS pain interference scales: results from three randomized clinical trials. Pain. 2018;159:775-82.

[35] Singh JA, Satele D, Pattabasavaiah S, Buckner JC, Sloan JA. Normative data and clinically significant effect sizes for single-item numerical linear analogue self-assessment (LASA) scales. Health Qual Life Outcomes. 2014;12:187.

[36] Butt Z, Wagner LI, Beaumont JL, Paice JA, Peterman AH, Shevrin D, et al. Use of a single-item screening tool to detect clinically significant fatigue, pain, distress, and anorexia in ambulatory cancer practice. J Pain Symptom Manage. 2008;35:20-30.

[37] Zimmerman M, Ruggero CJ, Chelminski I, Young D, Posternak MA, Friedman M, et al. Developing brief scales for use in clinical practice: the reliability and validity of single-item self-report measures of depression symptom severity, psychosocial impairment due to depression, and quality of life. J Clin Psychiatry. 2006;67:1536-41.

[38] Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain. 2010;150:173-82.

[39] Sloan JA, Aaronson N, Cappelleri JC, Fairclough DL, Varricchio C, Clinical Significance Consensus Meeting G. Assessing the clinical significance of single items relative to summated scores. Mayo Clin Proc. 2002;77:479-87.

[40] Deyo RA, Katrina R, Buckley DI, Michaels L, Kobus A, Eckstrom E, et al. Performance of a Patient Reported Outcomes Measurement Information System (PROMIS) Short Form in Older Adults with Chronic Musculoskeletal Pain. Pain Med. 2016;17:314-24.

[41] Beaumont JL, Fries JF, Curtis JR, Cella D, Yun H. Minimally important differences for Patient-Reported Outcomes Measurement Information System (PROMIS) fatigue and pain interference scores. Value Health. 2015;18:A165-A6.

[42] Amtmann D, Kim J, Chung H, Askew RL, Park R, Cook KF. Minimally important differences for Patient Reported Outcomes Measurement Information System pain interference for individuals with back pain. J Pain Res. 2016;9:251-5.

[43] Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. J Clin Epidemiol. 2011;64:507-16.

[44] Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res. 2007;16 Suppl 1:133-41.

[45] Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Qual Life Res. 2010;19:125-36.

[46] Varni JW, Magnus B, Stucky BD, Liu Y, Quinn H, Thissen D, et al. Psychometric properties of the PROMIS (R) pediatric scales: precision, stability, and comparison of different scoring and administration options. Qual Life Res. 2014;23:1233-43.

[47] Bingham CO, 3rd, Bartlett SJ, Merkel PA, Mielenz TJ, Pilkonis PA, Edmundson L, et al. Using patient-reported outcomes and PROMIS in research and clinical applications: experiences from the PCORI pilot projects. Qual Life Res. 2016;25:2109-16.

[48] Stucky BD, Huang W, Edelen MO. The Psychometric Performance of the PROMIS Smoking Assessment Toolkit: Comparisons of Real-Data Computer Adaptive Tests, Short Forms, and Mode of Administration. Nicotine Tob Res. 2016;18:361-5.

[49] Johns SA, Kroenke K, Theobald D, Wu J, Tu W. Telecare management of pain and depression in patients with cancer: patient satisfaction and predictors of use. J Ambulatory Care Management. 2011;34:126-39.

[50] Guilkey RE, Draucker CB, Wu J, Yu Z, Kroenke K. Acceptability of a telecare intervention for persistent musculoskeletal pain. J Telemed Telecare. 2018;24:44-50.

[51] Kean J, Monahan PO, Kroenke K, Wu J, Yu Z, Stump TE, et al. Comparative Responsiveness of the PROMIS Pain Interference Short Forms, Brief Pain Inventory, PEG, and SF-36 Bodily Pain Subscale. Med Care. 2016;54:414-21.

[52] Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61:102-9.

[53] King MT, Dueck AC, Revicki DA. Can methods developed for interpreting group-level patient-reported outcome data be applied to individual patient management? Med Care. 2019;57 Suppl 5 Suppl 1:S38-S45.

[54] Cella D, Choi S, Garcia S, Cook KF, Rosenbloom S, Lai JS, et al. Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. Qual Life Res. 2014;23:2651-61.

[55] Boyce MB, Browne JP. Does providing feedback on patient-reported outcomes to healthcare professionals result in better outcomes for patients? A systematic review. Qual Life Res. 2013;22:2265-78.

[56] Kroenke K, Evans E, Weitlauf S, McCalley S, Porter B, Williams T, et al. Comprehensive vs. Assisted Management of Mood and Pain Symptoms (CAMMPS) trial: Study design and sample characteristics. Contemp Clin Trials. 2018;64:179-87.

[57] Arroll B, Goodyear-Smith F, Kerse N, Fishman T, Gunn J. Effect of the addition of a "help" question to two screening questions on specificity for diagnosis of depression in general practice: diagnostic validity study. BMJ. 2005;331:884.

[58] Kroenke K, Krebs E, Wu J, Bair MJ, Damush T, Chumbler N, et al. Stepped Care to Optimize Pain Care Effectiveness (SCOPE) Trial: study design and sample characteristics. Contemp Clin Trials. 2013;34:270-81.

[59] Hoffman RM, Einstadter D, Kroenke K. Evaluating dizziness. Am J Med. 1999;107:468-78.

[60] Deyo RA, Rainville J, Kent DL. What can the history and physical examination tell us about low back pain? JAMA. 1992;268:760-5.

[61] Kroenke K. A practical and evidence-based approach to common symptoms: a narrative review. Ann Intern Med. 2014;161:579-86.

Table 1. Scale Scores for SPADE Symptoms at Baseline and 3-Months*

Scale[†]	Time point	N	Sleep	Pain	Anxiety	Depression	Fatigue
Numeric Rating Scale	Baseline	300					
Mean (SD)			6.1 (3.4)	5.9 (3.2)	5.0 (3.5)	4.5 (3.4)	5.9 (2.9)
Median (IQR)			7 (4-9)	7 (4-8)	5 (2-8)	5 (1-7)	6 (4-8)
PROMIS T-score	Baseline	300					
Mean (SD)			58.2 (9.0)	61.5 (9.4)	59.1 (9.4)	55.9 (9.8)	57.0 (10.0)
Median (IQR)			57.9 (52.4-63.8)	63.8 (55.6-67.9)	59.5 (53.7-65.3)	57.3 (49.0-62.2)	57.0 (51.0-64.6)
PROMIS T-score	3 Months	256					
Mean (SD)			53.6 (9.4)	59.1 (11.2)	56.4 (10.7)	53.3 (10.4)	54.1 (11.1)
Median (IQR)			54.3 (48.4-61.7)	61.2 (52.0-66.5)	57.7 (48.0-65.3)	55.7 (41.0-60.5)	53.1 (46.0-62.7)
Legacy Scale *(Possible Range)*	3 months	256	PSI-2 *(0-6)*	PEG *(0-10)*	GAD-2 *(0-6)*	PHQ-2 *(0-6)*	SF Vitality *(0-100)*
Mean (SD)			3.5 (1.8)	5.1 (3.3)	2.1 (2.1)	2.1 (2.0)	41.2 (22.3)
Median (IQR)			3 (2-5)	5.7 (2-8)	2 (0-4)	2 (0-4)	43.8 (25.0-56.3)

* Missing data was minimal. PROMIS T-scores were available for all 300 participants at baseline and in all 256 who completed follow-up questionnaires. Numeric rating scores were available for pain, fatigue and depression in 299 subjects, for anxiety in 298, and for sleep in 296. Legacy scale scores were available for fatigue and anxiety in all 256 participants who provided 3-month data, and for pain, sleep, and depression in 255.

[†] Possible range of scores for numeric rating scale is 0 to 10, with higher scores representing more severe symptoms. Range for PROMIS T-scores varies with the symptom scale but is 32 to 42 as the lowest possible score and 73 to 82 as the highest score, with 50 representing the population norm, and each 10 points representing one standard deviation; higher scores represent greater symptom severity.

Table 2. Operating Characteristics of NRS Cutpoints for Potential Clinical Symptom (PROMIS T-Score ≥ 55)

NRS Cutpoint *	Sensitivity	Specificity	Youden's Index	Positive Predictive Value	Negative Predictive Value
Pain					
≥ 4	.91	.68	.58	.91	.67
≥ 5	.83	.77	.60	.93	.56
≥ 6	.75	.89	.64	.96	.50
Anxiety					
≥ 4	.82	.76	.58	.90	.62
≥ 5	.75	.81	.56	.91	.55
≥ 6	.60	.89	.49	.94	.46
Fatigue					
≥ 4	.97	.44	.42	.74	.91
≥ 5	.93	.56	.48	.78	.82
≥ 6	.76	.75	.51	.83	.65
Sleep					
≥ 4	.97	.55	.52	.77	.93
≥ 5	.94	.64	.59	.80	.88
≥ 6	.88	.79	.66	.86	.81
Depression					
≥ 4	.79	.73	.52	.81	.70
≥ 5	.73	.82	.55	.85	.68
≥ 6	.61	.89	.50	.89	.61

***** The cutpoint of ≥ 5 is highlighted by gray shading as it represents a reasonable screening threshold.

Table 3. Correlations between PROMIS T-scores and Legacy Non-PROMIS Measures

Non-PROMIS measure	PROMIS T-Score *				
	Sleep	Pain	Anxiety	Depression	Fatigue
PSI-2 Sleep	**.83**	.57	.55	.53	.66
PEG Pain	.50	**.86**	.51	.46	.50
GAD-2 Anxiety	.55	.48	**.78**	.67	.52
PHQ-2 Depression	.48	.57	.68	**.78**	.62
SF-36 Vitality †	.56	.54	.52	.52	**.70**

***** Bold highlighting emphasizes highest correlations which are between same-symptom scales.

† Correlations actually negative since higher SF-36 scores represent better function but negative sign removed for simplicity of display in table.

Table 4. Cross-Walk between PROMIS T-scores and Legacy Non-PROMIS Measures

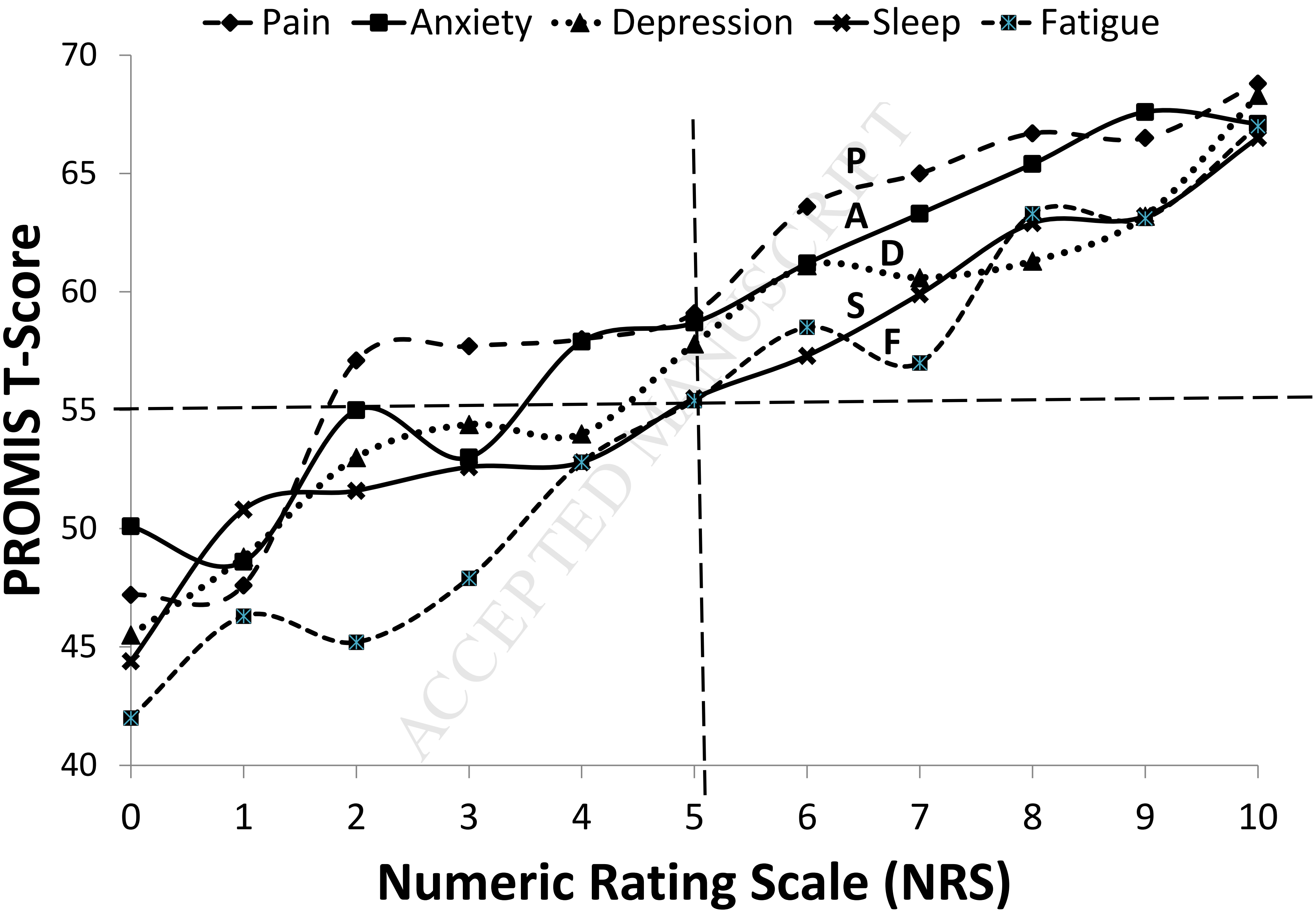
PROMIS Measure	**Legacy Measure (Range)**	**SD**	**α***	**Change in PROMIS T-score with:**				
				1-point change in legacy measure (95% CI)	0.20 SD change in legacy measure (95% CI)	0.35 SD change in legacy measure (95% CI)	1 SEM change[†] in legacy measure (95% CI)	0.50 SD change in legacy measure (95% CI)
Sleep	PSI-2 (0-6)	1.80	.837	4.36 (3.99 to 4.72)	1.57 (1.44 to 1.70)	2.74 (2.51 to 2.97)	3.16 (2.90 to 3.43)	3.92 (3.59 to 4.24)
Pain	PEG (0-10)	3.28	.947	2.92 (2.70 to 3.13)	1.91 (1.77 to 2.05)	3.35 (3.10 to 3.59)	2.20 (2.04 to 2.36)	4.78 (4.43 to 5.13)
Anxiety	GAD-2 (0-6)	2.08	.872	3.97 (3.57 to 4.37)	1.65 (1.48 to 1.81)	2.89 (2.60 to 3.18)	2.95 (2.65 to 3.25)	4.12 (3.71 to 4.54)
Depression	PHQ-2 (0-6)	1.96	.807	4.17 (3.76 to 4.58)	1.64 (1.47 to 1.80)	2.86 (2.58 to 3.15)	3.59 (3.24 to 3.95)	4.09 (3.69 to 4.49)
Fatigue	SF-36 Vitality (0-100)	22.3	.815	0.35 (0.30 to 0.39)	1.56 (1.37 to 1.76)	2.74 (2.40 to 3.08)	3.36 (2.94 to 3.78)	3.91 (3.42 to 4.40)
Mean MID estimate by 5 legacy scales					**1.67**	**2.92**	**3.05**	**4.16**

***** α = Cronbach's alpha

[†] SEM = standard error of measurement which is calculated as SD X square root (1 – Cronbach's alpha)

Legend (Caption)

Figure. Association between scores on single item 0 to 10 numeric rating scale (NRS) and mean PROMIS T-scores for the 5 SPADE symptoms.


Pain
Anxiety
Depression
Sleep
Fatigue
70
65
60
55
50
45
40
PROMIS T-Score
P
A
D
S
F
0
1
2
3
4
5
6
7
8
9
10
Numeric Rating Scale (NRS)

What is new?

Key findings

- Both single-item numeric rating scales (NRS) and 4-item PROMIS scales can efficiently assess the 5 SPADE symptoms (i.e., sleep problems, pain, anxiety, depression, and low energy/fatigue)
- A 3-point change in the PROMIS T-score represents a minimally important difference (MID), with 2 to 4 points representing the lower and upper MID bounds.

What this adds to what was known?

- An NRS score ≥ 5 is a useful screening cut-point for SPADE symptoms.

What is the implication and what should change now?

- Ultra-brief public domain scales can be used to detect and monitor common symptoms which are highly prevalent and burdensome in clinical practice.

All six authors have made substantial contributions to all of the following: (1) the conception and design of the study, or acquisition of data, or analysis and interpretation of data, (2) drafting the article or revising it critically for important intellectual content, (3) final approval of the version to be submitted.

CRediT author statement

Kurt Kroenke: Conceptualization, Methodology, Validation, Investigation, Writing - Original Draft, Visualization, Supervision, Funding Acquisition

Timothy Stump: Software, Formal Analysis, Data Curation, Writing - Review and Editing, Visualization

Jacob Kean: Conceptualization, Methodology, Validation, Investigation, Writing – Review and Editing

Tasneem Talib: Methodology, Validation, Investigation, Data Curation, Project Administration, Writing – Review and Editing

David Haggstrom: Conceptualization, Methodology, Validation, Investigation, Writing – Review and Editing

Patrick Monahan: Conceptualization, Methodology, Validation, Formal Analysis, Investigation, Writing – Reading and Editing, Visualization