Composite Measures of Pain, Anxiety, and Depressive (PAD) Symptoms:

Construct and Predictive Validity

M. A. Bushey, MD, PhD[1]*, K. Kroenke, MD[2,3,4],

F. Baye, MS[5]; S. Lourens, PhD[5]

[1] Department of Psychiatry, Indiana University School of Medicine, Indianapolis, IN

[2] VA HSR&D Center for Health Information and Communication, Roudebush VA Medical Center, Indianapolis, IN, United States

[3] Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, United States

[4] Regenstrief Institute, Inc., Indianapolis, IN, United States

[5] Department of Biostatistics, Indiana University Fairbanks School of Public Health, Indianapolis, IN, United States

* Corresponding author:

Michael A. Bushey, MD, PhD

Phone: 317-963-5517 | Fax: 317-963-7325 | E-mail: mabushey@iu.edu

Mailing Address: Goodman Hall, Rm 2800, 355 W. 16th Street, Indianapolis, IN 46202.

Running Title: Composite Measures of Pain, Anxiety, & Depression

Article data: 1 figure, 3 tables, word count = 3472, abstract length = 225 words

Conflicts of Interest: The authors declare they have no conflicts of interest to disclose.

_____

ABSTRACT

Objective: Pain, anxiety, and depression (PAD) are common, co-occurring symptoms that adversely affect one another and may respond to common treatments. PAD composite measures would be useful for tracking treatment response in patients with PAD symptoms. The goal of this study is to compare 3 different PAD composite scales in terms of construct validity, responsiveness, and utility in predicting global improvement.

Method: The sample consisted of 294 primary care patients enrolled in a telecare trial for treating pain, anxiety, and depression. Assessments at baseline and 3 months included the Brief Pain Inventory, PHQ-9 depression scale, GAD-7 anxiety scale, PROMIS measures, Medical Outcomes Study Short-Form items, disability measures, and patient-reported global improvement. Construct validity of the PAD composite measures, their responsiveness, and their ability to predict global improvement was analyzed using Pearson correlations, standardized response means, and receiver operating characteristics analysis.

Results: PAD composite measures correlated strongly with one another, and moderately with measures of function, vitality, and disability. Each PAD composite measure demonstrated similar responsiveness in detecting improvement at 3 months as assessed by standardized response means (SRMs) and area under the curve (AUC analyses).The SRMs for partial and substantial global improvement corresponded to moderate (Cohen's d of .58 to .69) and large (.81 to .93) effect sizes, respectively.

Conclusions: Three different PAD composite measures demonstrate good construct validity as well as responsiveness in detecting global improvement of pain, anxiety and depression at 3 months.

Key Words: Pain, Anxiety, Depression, Composite Measures, PROMIS measures, Psychometrics

1. INTRODUCTION

Pain is the most prominent physical complaint in the general population and is the leading symptom driving primary care visits [1-3]. Depression and anxiety are the most prevalent mental health symptoms [1, 2, 4], and are known to frequently co-occur with pain, leading to significant disability and reciprocal adverse effects [5-7]. Depression and anxiety can amplify negative experiences such as pain, while pain can in turn lead to decreases in mood, and pain-related activity avoidance can contribute significantly to anxiety [8]. The estimated prevalence of depression (2-83%) and anxiety (1-65%) in patients with chronic pain varies widely depending on the population, pain condition, and method of determination [9]. A cross-sectional study of 2981 patients in the Netherlands found that pain severity and disability increased as both depression and anxiety severity increased [10]. Because of the importance of this pain-anxiety-depression (PAD) triad and the interdependence of these symptoms, comprehensive treatment involves targeting all symptoms simultaneously.

In clinical care and research trials, PAD symptoms are commonly assessed using dedicated patient reported outcome measures (PROMs). Measures such as the Patient Health Questionnaire 9-item depression scale (PHQ-9) [11], Generalized Anxiety Disorder 7-item scale (GAD-7) [12], and the Brief Pain Inventory (BPI) 11-item scale [13], have been in widespread use in clinical care and research studies for over a decade. PAD symptoms can also be assessed as components of multidimensional scales, such as the Medical Outcomes Study Short Form 36 (SF-36). the most widely used assessment of health-related quality of life [14, 15] Another multidimensional measure, the Patient-Reported Outcomes Measurement Information System (PROMIS) was developed by the National Institute of Health to provide a well-validated, publicly-available bank of PROMs to assess a comprehensive set of domains, including the PAD triad symptoms [16, 17].

Whereas dedicated measures assess a single construct, and multidimensional measures multiple constructs, composite measures, which combine individual domain assessments into a single composite score, are attractive for several reasons.When assessing PAD symptoms, a composite score would provide a single variable for quantifying the collective burden of disease – as is the goal with global health measures – while also retaining underlying symptom-level scores. Since numerous pharmacologic and non-pharmacologic treatments target all three PAD symptoms simultaneously [18], a composite measure would provide a convenient outcome measure in clinical trials and practice. Composite scales utilizing PROMIS measures have previously demonstrated efficacy in measuring the SPADE clinical pentad (**s**leep, **p**ain, **a**nxiety, **d**epression, low **e**nergy/fatigue) [19, 20]. Likewise, composite scales have been helpful in assessing symptom dyads, such as depression and anxiety [21], as well as two dimensions of a single symptom (e.g., pain intensity and pain interference) [22].

The Comprehensive vs. Assisted Management of Mood and Pain Symptoms (CAMMPS) clinical trial compared two levels of telecare treatment targeting pain, depression, and anxiety symptoms. The primary outcome measure of this trial was a PAD composite score that combined the PHQ-9, GAD-7, and BPI into a single z-score. We found the composite z-score to be a highly pragmatic and helpful measure of overall progress in patients with PAD triad symptoms, and can be a useful target for research studies treating this common symptom cluster or clinical systems that already utilize some or all of the component measures. In this paper, we analyze data from the CAMMPS trial to more fully characterize the psychometric properties of the PAD composite z-score. Given that the z-score composite does require calculation, we also compare it to two additional PAD composite measures that can be more easily computed in busy clinics: one that utilizes publicly-available and highly standardized PROMIS measures, and one derived from the proprietary but widely used and studied Medical Outcome Study Short Form (SF) scales. We hypothesized that these 3 PAD composite scales would be comparable to each other in terms of construct validity and their performance in

predicting clinical global improvement. Our comparison of these 3 approaches for measuring the PAD triad may be useful to researchers or clinical enterprises looking to develop measurement-based care.

2. METHODS

2.1 Study Participants

Data for this study was collected as part of the CAMMPS trial which is described in detail elsewhere [2]. All procedures performed in the CAMMPS trial were reviewed and approved by the Indiana University Institutional Review Board and the Veteran's Administration (VA) Research Review Committee and carried out in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study. Briefly, patients with pain and mood symptoms were recruited from five primary care clinics in a Midwestern VA Medical Center. Participants were randomized to one of two different telecare interventions for pain and mood symptoms in a 12-month trial. The lower intensity treatment arm received automated symptom monitoring and a web-based self-management program (ASM group), while the higher intensity arm received these interventions plus treatment monitoring and adjustment by a nurse care manager-physician team (CSM group).

Eligibility criteria included age ≥ 18 years old and clinically significant musculoskeletal pain plus comorbid depression and/or anxiety. Pain had to be present at least 3 months despite treatment with analgesics and moderate in severity as assessed by the Brief Pain Inventory Scale (BPI). Moderate severity depression and/or anxiety was defined as either: (1) a PHQ-8 score ≥ 10 with depressed mood and/or anhedonia endorsed; (2) GAD-7 score ≥ 10; or (3) a combined GAD-7 and PHQ-8 score ≥ 12. The PHQ-8 is a variant of the PHQ-9 commonly used for screening that is identical to the original measure with the exception that it omits the item about self-harm yet has scoring thresholds identical to the PHQ-9 [23]. Exclusion criteria

included: not speaking English; significant cognitive impairment; severe mental illness not suitable for telecare including suicidality, schizophrenia, or bipolar disorder; pregnancy; or a life expectancy less than 12 months.

2.2 Study Measures

2.2.1 Measure Administration

Measures assessing pain, anxiety, depression, disability, health-related quality of life, medical co-morbidity, and global impression of change were collected at baseline, 1, 3, 6, and 12 months by trained interviewers blinded to treatment arm [2]. For the analysis presented here, baseline and 3-month data were examined. Measures were administered by a research interviewer either in person (baseline) or via telephone (follow-up).

2.2.2 PAD composite z-score

A composite z-score was calculated from the BPI, GAD-7, and PHQ-9. The BPI is an 11-item instrument that assesses pain severity (4 items) and pain interference (7 items) [13]. Each item is scored from 0 to 10 with higher scores representing worse pain, and a total BPI score is generated by averaging the 11 individual item scores. The GAD-7 is a 7-item instrument that assesses the intensity of anxiety experienced over the past two weeks, with overall scores ranging from 0 to 21. Higher scores reflect greater severity of anxiety [12]. The PHQ-9 includes 9 items that assesses the severity of depression experienced over the past two weeks [11]. Overall scores range from 0 to 27, with higher scores indicating greater depression severity.

To calculate the composite z-score, we first calculated z-scores for the individual BPI, GAD7, and PHQ9 scores [1]. The z-score is the person's scale score minus the sample mean at baseline divided by the baseline sample standard deviation. The composite PAD z-score is the average of the z-scores for the 3 scales. The change in z-score over time represents the number of standard deviations by which an individual's PAD symptoms improve or worsen over the baseline sample mean. Because the 3 scales have different response options and score

ranges, conversion to z-scores allowed us to standardize scores across the 3 scales and derive a composite score. The change in z-score is equivalent to an effect size for which 0.2, 0.5, and 0.8 represent small, moderate, and large clinical changes, respectively [24].

2.2.3 PROMIS Measures

PROMIS includes an array of measures that can evaluate multiple domains of physical and mental health [16, 25]. PROMIS scales ask patients to indicate the severity of symptoms over the past 7 days with 5 response options. By totaling the responses, a raw score is generated for each PROMIS scale. The 8-item scales for pain, anxiety and depression from the PROMIS-57 profile were collected. Within the 8-item scale are nested 6 and 4 item scales. Raw scores range from 8 to 40 on the 8-item scales, 6 to 30 on the 6-item scales, and 4-20 on the 4-item scales. T-scores are generated from the raw scores with a general population mean of 50 and a standard deviation of 10; higher scores reflect more severe symptoms. The PROMIS PAD-composite raw score was created by adding together the individual raw scores of the pain, anxiety, and depression subscales and dividing the total by 3. Likewise, the PROMIS PAD-composite T-score was derived by averaging the T-scores of the individual subscales.

2.2.4 Other Measures

We assessed health-related quality of life with the Medical Outcomes Study Short-Form 12-item scale (SF-12) [26] plus 8 items from the SF-36, allowing assessment of the SF-36 bodily pain, social functioning, vitality, mental health, and current general health scores [14]. The bodily pain subscale consists of 1 pain severity and 1 pain interference item, and the mental health subscale consists of 3 depression and 2 anxiety items. Scores on each SF subscale range from 0 to 100 with lower scores representing worse health-related quality of life. The SF PAD was calculated from these measures by averaging the bodily pain and mental health subscale scores.

Disability was measured with three different instruments. The Sheehan Disability Scale (SDS) assesses the extent that health has interfered with family life, social life, and work over the past month on a scale of 0 to 10 with higher scores reflecting greater disability [27]. The overall SDS score is the mean of these three items. Disability days were quantified with a single question that asked participants to report the number of days during the preceding 4 weeks that they either had to reduce usual activities by at least 50% due to physical or mental health problems or were in bed [2]. Work effectiveness was qualified with a single question asking how effective on his or her job the participant was during the past 4 weeks on a scale of 0% to 100%, with higher scores reflecting better work effectiveness.

Global impression of change was measured with two items, one assessing change in pain from the start of the trial, and the other assessing change in low mood or anxiety [2]. Respondents were asked to rate their symptoms as 1 (much better), 2 (moderately better), 3 (a little better), 4 (no change), 5 (a little worse), 6 (moderately worse), or 7 (much worse). A composite global impression of change score was created by summing the responses to the pain and mood/anxiety questions, with a range of 2 to 14. A categorical global composite score was also calculated for which responses to each individual question were collapsed into 0 = worse (for responses 5, 6 and 7), 1 = same (for response 4), and 2 = improved (for responses 1, 2 or 3). These categorical scores for the pain and mood/anxiety items were summed to create the global categorical change score, which ranged from 0 to 4. A score of either "0" or "1" was categorized as overall worsening, as it reflected worsening of both symptoms or worsening of one symptom with no change in the other. A score of "4" reflected improvement in both pain and mood and was considered globally "much improved". A score of "3" required improvement in one symptom and no change in the other and was categorized as "partially improved". A score of "2" was categorized as globally being "the same", and could be achieved by having both pain and mood scored as "no change", or improvement in one symptom being cancelled out by worsening of the other symptom.

2.3 Statistical Analysis

SAS Version 9.3 was used for all analyses. There was a range of global change in both groups with improvement, no change, and worsening occurring in 55%, 26%, and 19% of participants in the CSM arm vs. 44%, 28% and 28% of participants in the ASM arm [1]. In evaluating a scale's responsiveness, a reasonable distribution of change in the sample is what is important rather than which factors caused the change. Thus, for this psychometric paper, study participants were analyzed as a combined cohort rather than by treatment arm.

To assess for construct validity, baseline PROMIS PAD raw and T-scores were correlated with the PAD z-score, SF-PAD, SDS, disability days, work effectiveness, SF social functioning, SF vitality, and SF general health. Responsiveness of the composite PAD scores was evaluated using several metrics [22]. First, we examined the average absolute change for the PAD scores from baseline to 3 months for patients who fell into each category of global change at 3 months. Second, because each composite PAD scale used different units, standardized response means were also calculated for the change scores (except for the PAD z-score, which is already a standardized score) by dividing the 3-month change score by the SD of the within-group change score. Third, correlations between global improvement and the baseline to 3-month change scores for PROMIS PAD, SF PAD, and PAD z-score were calculated.

Fourth, the ability of the PAD composite scores to detect global improvement was investigated using area under the curve (AUC) values determined by receiver operating characteristic (ROC) analyses. Assessing responsiveness to change is analogous to assessing the discriminatory ability of a diagnostic test; therefore, ROC curves can be used to assess a measure's ability to accurately "diagnose" the presence or absence of a clinically important change [22,28]. ROC curves plot sensitivity on the y-axis against (1 - specificity) of the x-axis for a measure compared with a reference standard (in this case, global improvement). AUC values range from 0.5 (same as chance) to 1.0 (complete discrimination).

3. RESULTS

3.1 Study Participants

A total of 294 participants were enrolled in the CAMMPS study, with an overall mean age of 57.4 years (range 25 to 88). The sample population was 87.4% male, 79.3% white, 15.3% black, and 5.4% other race. 21% of the sample were college graduates, while 53% had completed at least some college or trade school, and 26% had a high school equivalent as their highest level of education. Randomization succeeded in making both arms of the trial similar in terms of patient characteristics [1, 2], with the only difference between groups being the rate of probable post-traumatic stress disorder (PTSD) (49.7% in CSM group vs. 61.2% in ASM group, P = 0.047) For the analyses in this paper, Therefore, participants from both arms of the trial were combined for PAD analyses.

3.2 Global Improvement

At 3 months, follow-up data was available on 259 patients, of whom 59 (22.7%) reported worse pain, 105 (40.5%) reported the same pain, while 95 patients (36.7%) reported improved pain. A similar pattern was observed for mood symptoms, with 48 patients (18.5%) reporting worse mood or anxiety, 100 patients (38.6%) reporting the same mood or anxiety, and 111 patients (42.9%) reported improved mood. The Global Improvement in pain and mood, calculated by summing together the full 7-item responses for the pain and mood improvement questions, had a symmetrical distribution (mean = 7.4, median = 7, mode = 8, skewness = 0.13, kurtosis = 0.48). (Figure 1a). When the Global Improvement is viewed categorically by collapsing the pain and mood responses into worse, same, or better groups, the overall total who are overall worse (0 or 1) is 61 patients (23.6%), those who are "the same" (2) is 69 (26.6%), and those who are overall improved (3 or 4) is 129 (49.8%). Figure 1b shows the distribution of the individual scores that contribute to these composite improvement scores.
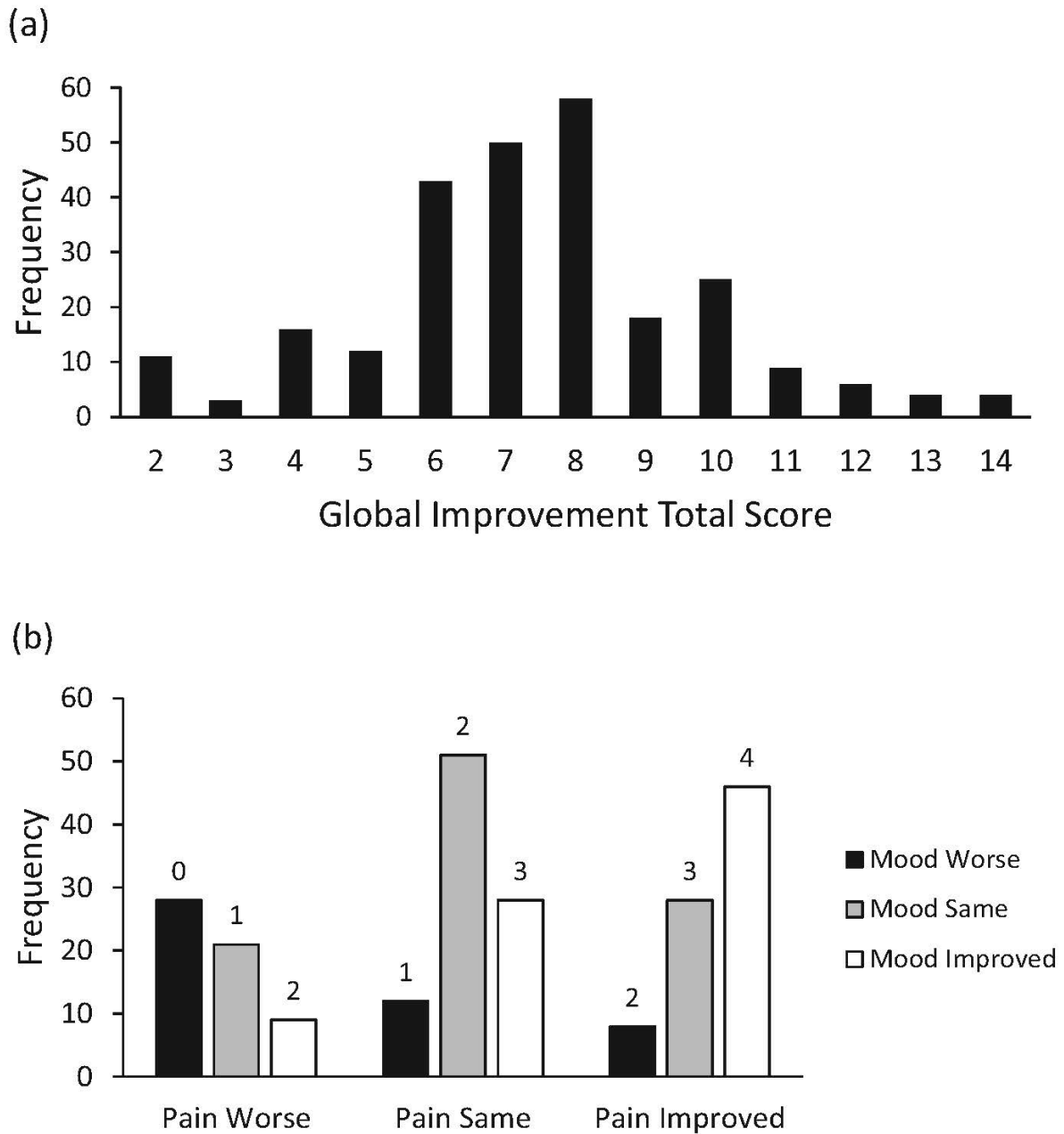
Fig 1 a. Frequency Distribution of Global Improvement Total Score, calculated using the full 7 point response for each item (pain + mood). **b.** Breakdown of Frequency Distribution of Global Mood and Global Pain Changes. Data Labels show the Global Improvement Categorical value for each grouping.

3.3 Construct validity of composite PAD measures

Table 1 demonstrates strong convergent validity among the PAD composite measures. High correlations (Pearson correlations ranged 0.74 – 0.86) were observed between the PROMIS PAD scores, PAD z-score, and the SF-PAD. Performance of the PROMIS PAD measures were similar among the 4, 6, and 8-item PAD measures and did not differ between the raw and converted T scores. Specifically, Pearson correlations differed by no more than 0.04 between different versions of the PROMIS PAD and other measures analyzed. All PAD composite scores correlated moderately (Pearson correlations ranged 0.52 – 0.72) with measures of function including the SDS, disability days, SF social functioning, and SF vitality score. Correlations were slightly lower (0.33 – 0.43) for work effectiveness and SF general health. Not listed in the table are the correlations between the six different permutations of PROMIS PAD composites (4, 6 and 8-item raw scores and T-scores) which were nearly perfectly correlated (0.96 – 0.99) as expected given their substantial item overlap.

Table 1 Correlation of PAD Composite Scores and Measures of Construct Validity at Baseline (n=294)

	PROMIS PAD Raw Score			PROMIS PAD T-score			Other PAD	
Measure	**4-item**	**6-item**	**8-item**	**4-item**	**6-item**	**8-item**	**z-score**	**SF**
z-score PAD [a]	0.77	0.78	0.79	0.75	0.76	0.76	--	-0.76
Short-Form (SF) PAD	-0.81	-0.82	-0.83	-0.80	-0.81	-0.81	-0.76	--
Sheehan Disability	0.70	0.70	0.71	0.67	0.67	0.67	0.69	-0.70
Disability Days	0.57	0.57	0.57	0.55	0.54	0.54	0.52	-0.57
Work Effectiveness	-0.37	-0.38	-0.38	-0.34	-0.35	-0.35	-0.39	0.40
SF Social Function	-0.69	-0.70	-0.71	-0.66	-0.67	-0.67	-0.64	0.64
SF Vitality	-0.58	-0.59	-0.59	-0.57	-0.58	-0.57	-0.58	0.62
SF General Health	-0.40	-0.40	-0.40	-0.37	-0.38	-0.37	-0.36	0.41

[a] z-score PAD = composite of BPI, PHQ-9, and GAD-7 z-scores

3.4 Comparative Responsiveness of Composite PAD Measures

To assess the responsiveness of the PAD composite measures, we calculated the change in each PAD composite score from baseline to 3 months. The average change score for each measure was calculated for patients categorized as either globally worse, the same, partially improved, or much improved as indicated by their PAD global categorical score. As shown in Table 2, each PAD composite measure distinguished between the levels of global change. Patients who were either partially or much improved according to their global categorical scores had correspondingly large improvements in their PAD composite scores.

Table 2 Comparative Responsiveness of Composite PAD Measures (n = 259)

	Change Score Mean (SD) by Global Change Category [a]			
Composite PAD measure	0, 1 Worse (n=61)	2 Same (n=69)	3 Partially improved (n=70)	4 Much improved (n=59)
PROMIS 4-item	0.49 (5.18)	0.74 (4.91)	**3.08 (4.46)**	**5.06 (5.53)**
PROMIS 6-item	0.47 (5.38)	0.78 (4.95)	**3.14 (4.57)**	**4.78 (5.77)**
PROMIS 8-item	0.52 (5.21)	0.83 (4.97)	**2.98 (4.80)**	**5.03 (5.43)**
SF-PAD	1.52 (10.23)	-0.54 (11.23)	**-6.21 (10.71)**	**-11.83 (14.58)**
PAD z-score	0.08 (0.67)	0.25 (0.74)	**0.62 (0.68)**	**0.83 (0.90)**

[a] Change Score= (baseline – 3 month). A positive change score represents improvement for all scales except the SF for which a negative score represents improvement.

Bold values- indicated significant difference ($p < .05$) in change scores between reference (i.e. worse) and the other groups (There was no significant difference between 'worse' and 'same' group for any of the PAD measures)

In order to compare across the PAD composites, change scores for the PROMIS PAD and SF-PAD composite were converted into standardized response means (SRMs). As noted, this step was unnecessary for the PAD z-score, which is already a standardized score. Table 3

shows the SRM comparisons, which demonstrate a moderate effect size for all PAD composite

scores in the partially improved category, and a large effect size in the much improved category.

Correlations between the PAD composite change scores and global change continuous score

ranged from 0.33 – 0.44. The PAD composite measures were comparable in detecting global

improvement as indicated by a relatively similar area under the curve (AUC): PROMIS 4-item

AUC = 0.678 (SE = 0.033), PRMOIS 6-item AUC = 0.669 (SE = 0.033), PROMIS 8-item AUC =

0.669 (SE = 0.033), Short Form (SF) AUC = 0.708 (SE 0.032), and z-score AUC = 0.696 (SE =

0.033).

Table 3 Standardized Response Mean (SRM) Comparison of Composite PAD Measures

Composite PAD measure	SRM (95% CI) by Global Change Category			
	0, 1 Worse (n=61)	2 Same (n=69)	3 Partially improved (n=70)	4 Much improved (n=59)
PROMIS 4-item	0.10 (-0.16, 0.35)	0.15 (-0.09, 0.39)	0.69 (0.45, 0.93)	0.91 (0.65, 1.18)
PROMIS 6-item	0.09 (-0.17, 0.34)	0.16 (-0.08, 0.40)	0.69 (0.45, 0.92)	0.83 (0.57, 1.09)
PROMIS 8-item	0.10 (-0.16, 0.36)	0.17 (-0.07, 0.41)	0.62 (0.38, 0.86)	0.93 (0.67, 1.19)
Short Form (SF)	0.15 (-0.11, 0.40)	-0.05 (-0.29, 0.19)	-0.58 (-0.82, -0.34)	-0.81 (-1.07, -0.55)
z-score [a]	0.08 (-0.09, 0.26)	0.25 (0.07, 0.43)	0.62 (0.46, 0.78)	0.83 (0.59, 1.06)

[a] SRM cannot be calculated for z-score since the latter is already standardized. However, the z-score is comparable to SRM so it is included in the table in order to compare its magnitude with SRMs for the other measures.

4. DISCUSSION

Our study has several important findings. First, we found that three different PAD composite measures (PAD z-score, PAD PROMIS, and SF PAD) highly correlated with each other and had good construct validity. Second, we found that changes in the PAD composite scores at 3 months were predictive of patient-reported global improvement. Third, AUC analyses demonstrated that the three types of PAD composite measures were similar in their ability to detect global improvement.

In our recent CAMMPS clinical trial evaluating the comparative efficacy of two telecare approaches to treat pain, depression, and anxiety, we developed a PAD composite z-score that combined results of the PHQ-9, GAD-7, and BPI-interference into a single outcome measure. This composite measure was comprised of well-validated measures and proved sensitive to change in detecting intervention effects. The PHQ-9/GAD-7/BPI z-score method may be attractive in health systems that are already collecting these measures as part of routine care. This would require knowing the population mean and standard deviation, and calculating the z-scores, which could potentially be automated in electronic record or case management tracking tools. However, the degree of complexity involved with this calculation, along with the length of the three measures (27 questions in total), may impede widespread implementation of this approach. Further, additional measures assessing sleep and fatigue would be required if capturing data on the full SPADE pentad would be desired in the future.

Alternatively, if organization choice is not constrained by legacy scales in use, implementing the PROMIS scales would be an attractive option, with the benefit of having ready-to-go sleep and fatigue measures that could be added. Using the PROMIS 4-item PAD composite (12 questions total) would reduce response burden compared to the legacy z-score approach. It would also be easy to add in sleep and fatigue PROMIS measures to allow for assessment of the full SPADE pentad [28]. For organizations or research studies that are already using or planning to use the SF-36 (or the SF-12 + 8 approach as utilized in this study) to measure health-related quality of life, calculating the SF-PAD would allow the generation of a

useful PAD composite metric without administering additional measures. The SF-36 does include a fatigue item, though an additional scale would need to be added to assess the full SPADE pentad..

Importantly, the three PAD composites analyzed demonstrated similar construct validity and responsiveness. As shown in Table 2, patients who were categorized as partially improved by global change averaged an improvement of approximately 3 points in their PROMIS scores, which corresponds to previously calculated minimally-important differences (MID) for PROMIS measures of pain, anxiety and depression symptoms [20]. All three measures demonstrated similar sensitivity to change as shown by comparison of SRMs in Table 3. Patients reporting partial improvement had SRMs representing a moderate effect size (0.58 to 0.69), and those reporting much improvement had SRMs representing a large effect size (0.81 to 0.93). Of note, the PAD composites were better at showing improvement over time than worsening, which is consistent with previous studies of patient-reported outcomes [22, 29, 30]. It is possible that patients whose baseline pain and mood scores are already quite elevated may reach a "ceiling" preventing their scores from worsening despite a subjective feeling of being worse. Another reason might be that patients are less able to discriminate no change from worsening of symptoms since, in both instances, the outcome is suboptimal and may necessitate further treatment changes.

In addition to comparing three distinct PAD composites (PAD z-score, PAD PROMIS, and SF PAD), we also compared the PROMIS PAD 8-item scores to the shorter 4 and 6 item versions of the scales and the raw scores to the T-scores. We found similar construct validity and responsiveness regardless of PROMIS scale length or type of score. This suggests that, in busy clinic settings, shorter scales may be adequate and that using the raw PROMIS PAD scores may be sufficient if the computerized program needed to convert raw to T-scores is not available.

Our study has several strengths. Data was gathered longitudinally as part of a controlled trial targeting the PAD symptoms. In addition, there were high rates of assessment completion among participants. Outcomes included a well-validated set of measures, allowing comparison of three distinct PAD composite scales. The study also collected patient global impression of change (PGIC) data at each time point as an independent anchor for defining improvement or worsening of symptoms.

Our study was limited due to its relatively homogenous sample, a consequence of collecting data in a clinical trial that recruited predominantly male veterans with chronic and comorbid mental health problems. Future studies in patients without chronic pain, non-veteran samples with more females, and in settings other than primary care will be required to better assess generalizability. Also, the AUCs demonstrate moderate rather than strong differentiation between subjects whose symptoms had improved and those who had not. This AUC range is consistent with other studies using PGIC as an anchor [22, 29-31]. In these studies, the similarity of AUCs when using PGIC as a reference in receiver operating characteristics curves is more useful in determining the similarity of different tests than in assessing their absolute value.

Our findings demonstrate that composite measures can be a valid way of assessing and following treatment response of PAD symptoms over time. Further, three different composite scales demonstrated similar utility and predictive value in assessing global improvement of pain and mood symptoms. The PROMIS measures are public domain and scored on the same scale making it easy to create a composite score without using complex statistical tools. Thus, the PROMIS PAD measure would likely be the most feasible to implement for clinics interested in deploying a composite PAD metric. The SF PAD consists of items from the SF-36 which in turn is a proprietary scale meaning that there may be a fee for its use. On the other hand, at 7 items the SF PAD is the shortest composite measure and also lends itself to simple calculation of a composite score (i.e., the average of the SF-36 bodily pain and mental health scales). In

research studies or clinical practice settings where the SF-36 is already being used, the SF PAD is an attractive alternative. In summary, monitoring PAD symptoms both individually and collectively in research and practice is feasible by the use of brief scales which can provide scores specific to each symptom as well as a composite score representing the combined effect of the individual symptoms.

Funding

Disclosures

Conflicts of interest: none

Author Contributions

M. A. Bushey: data analysis and interpretation, discussion of results, manuscript drafting and revision, final approval of the version to be published. K. Kroenke: design of the study, data analysis and interpretation, discussion of results, manuscript drafting and revision, final approval of the version to be published. F. Baye: data analysis and interpretation, discussion of results, final approval of the version to be published. S. Lourens: design of the study, data analysis and interpretation, discussion of results, final approval of the version to be published.

REFERENCES

1.   Kroenke, K., et al., *Automated self-management (ASM) vs. ASM-enhanced collaborative care for chronic pain and mood symptoms: the CAMMPS randomized clinical trial.* J Gen Intern Med, 2019.

2.   Kroenke, K., et al., *Comprehensive vs. Assisted Management of Mood and Pain Symptoms (CAMMPS) trial: study design and sample characteristics.* Contemporary clinical trials, 2018. **64**: p. 179-187.

3.   Institute of Medicine, *Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education, and Research.* 2011, Washington, DC: National Academies Press.

4.   Kroenke, K. and J. Unutzer, *Closing the false divide: sustainable approaches to integrating mental health services into primary care.* Journal of general internal medicine, 2017. **32**: p. 404-410.

5.   Kroenke, K., et al., *Reciprocal relationship between pain and depression: a 12-month longitudinal analysis in primary care.* J Pain, 2011. **12**(9): p. 964-73.

6.   Murray, C.J.L., et al., *The State of US Health, 1990-2010 Burden of Diseases, Injuries, and Risk Factors.* Jama-Journal of the American Medical Association, 2013. **310**(6): p. 591-608.

7.   Scott, E.L., et al., *Beneficial effects of improvement in depression, pain catastrophizing, and anxiety on pain outcomes: a 12-month longitudinal analysis.* J Pain, 2016. **17**(2): p. 215-22.

8.   Peters, M.L., *Emotional and Cognitive Influences on Pain Experience.* Mod Trends Pharmacopsychiatry, 2015. **30**: p. 138-52.

9.   Hooten, W.M., *Chronic Pain and Mental Health Disorders: Shared Neural Mechanisms, Epidemiology, and Treatment.* Mayo Clin Proc, 2016. **91**(7): p. 955-70.

10.   de Heer, E.W., et al., *The association of depression and anxiety with pain: a study from NESDA.* PLoS One, 2014. **9**(10): p. e106907.

11.   Kroenke, K., R.L. Spitzer, and J.B. Williams, *The PHQ-9: validity of a brief depression severity measure.* J Gen Intern Med, 2001. **16**(9): p. 606-13.

12.   Lowe, B., et al., *Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population.* Med Care, 2008. **46**(3): p. 266-74.

13.   Cleeland, C.S., et al., *Pain and its treatment in outpatients with metastatic cancer.* N Engl J Med, 1994. **330**(9): p. 592-6.

14.   Ware, J.E., Jr. and C.D. Sherbourne, *The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection.* Med Care, 1992. **30**(6): p. 473-83.

15.   Pequeno, N.P.F., et al., *Quality of life assessment instruments for adults: a systematic review of population-based studies.* Health Qual Life Outcomes, 2020. **18**(1): p. 208.

16.   Cella, D., 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**(11): p. 1179-94.

17.   Gershon, R.C., et al., *The use of PROMIS and assessment center to deliver patient-reported outcome measures in clinical research.* J Appl Meas, 2010. **11**(3): p. 304-14.

18.   Kroenke, K., *A practical and evidence-based approach to common symptoms: a narrative review.* Ann Intern Med, 2014. **161**(8): p. 579-86.

19.   Davis, L.L., et al., *The SPADE symptom cluster in primary care patients with chronic pain.* Clin J Pain, 2016. **32**(5): p. 388-93.

20.   Kroenke, K., et al., *PROMIS 4-item measures and numeric rating scales efficiently assess SPADE symptoms compared with legacy measures.* J Clin Epidemiol, 2019. **115**: p. 116-124.

21.   Kroenke, K., et al., *Patient Health Questionnaire Anxiety and Depression Scale: initial validation in three clinical trials.* Psychosom Med, 2016. **78**(6): p. 716-27.

22. Kean, J., 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**(4): p. 414-21.

23. Wu, Y., et al., *Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis.* Psychol Med, 2019: p. 1-13.

24. Kazis, L.E., J.J. Anderson, and R.F. Meenan, *Effect sizes for interpreting changes in health status.* Med Care, 1989. **27**(3 Suppl): p. S178-89.

25. Pilkonis, P.A., et al., *Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS(R)): depression, anxiety, and anger.* Assessment, 2011. **18**(3): p. 263-83.

26. Ware, J., Jr., M. Kosinski, and S.D. Keller, *A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity.* Med Care, 1996. **34**(3): p. 220-33.

27. Sheehan, D.V., K. Harnett-Sheehan, and B.A. Raj, *The measurement of disability.* Int Clin Psychopharmacol, 1996. **11 Suppl 3**: p. 89-95.

28. Kroenke, K., et al., *Incorporating PROMIS Symptom Measures into Primary Care Practice-a Randomized Clinical Trial.* J Gen Intern Med, 2018. **33**(8): p. 1245-1252.

29. Krebs, E.E., et al., *Comparative responsiveness of pain outcome measures among primary care patients with musculoskeletal pain.* Med Care, 2010. **48**(11): p. 1007-14.

30. Kroenke, K., et al., *Comparative responsiveness of pain measures in cancer patients.* J Pain, 2012. **13**(8): p. 764-72.

31. Johns, S.A., et al., *Longitudinal comparison of three depression measures in adult cancer patients.* J Pain Symptom Manage, 2013. **45**(1): p. 71-82.