Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

HHS Public Access
Author manuscript
Med Care. Author manuscript; available in PMC 2016 May 22.

Published in final edited form as:
Med Care. 2010 November ; 48(11): 1007–1014. doi:10.1097/MLR.0b013e3181eaf835.

Comparative Responsiveness of Pain Outcome Measures Among Primary Care Patients With Musculoskeletal Pain

Erin E. Krebs, MD, MPH[*,†,‡], **Matthew J. Bair, MD, MS**[*,†,‡], **Teresa M. Damush, PhD**[*,†,‡], **Wanzhu Tu, PhD**[†,‡], **Jingwei Wu, MS**[‡], and **Kurt Kroenke, MD**[*,†,‡]

[*]Center on Implementing Evidence-Based Practice, Roudebush VA Medical Center, Indianapolis, IN

[†]Department of Medicine, Indiana University School of Medicine, Indianapolis, IN

[‡]Regenstrief Institute, Inc., Indianapolis, IN

Abstract

Background—Comparative responsiveness data are needed to inform choices about pain outcome measures.

Objectives—To compare responsiveness of pain intensity, pain-related function, and composite measures, using data from a randomized trial and observational study.

Research Design—Analysis of responsiveness.

Subjects—A total of 427 adults with persistent back, hip, or knee pain were recruited from primary care.

Methods—Participants completed Brief Pain Inventory, Chronic Pain Grade (CPG), Roland disability, SF-36 bodily pain, and pain global rating of change measures. We used the global rating as the anchor for standardized response mean and receiver operating characteristic curve analyses. We used the distribution-based standard error of measurement to estimate minimally important change. To assess responsiveness to the trial intervention, we evaluated standardized effect size statistics stratified by trial arm.

Results—All measures were responsive to global improvement and all had fair-to-good accuracy in discriminating between participants with and without improvement. SF bodily pain was less responsive than other measures in several analyses. The 3-item PEG was similarly responsive to full Brief Pain Inventory scales. CPG and SF bodily pain were less responsive to the trial intervention and did not perform well among participants with hip/knee pain. Agreement between anchor and distribution-based methods was modest.

Conclusions—If a brief measure is desired, the 3-item PEG is more responsive than the SF bodily pain scale. CPG and SF bodily pain scales may be relatively poor choices for trial outcome assessment. Both anchor and distribution-based methods should be considered when determining clinically important change.

Reprints: Erin E. Krebs, MD, MPH, Roudebush VAMC (11H), 1481 W. 10th St, Indianapolis, IN 46202. ; Email: krebse@iupui.edu.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Keywords

pain; measurement; psychometrics; responsiveness

Chronic pain research depends on valid patient-reported outcome measures to evaluate symptom severity, understand disease prognosis, and assess treatment effectiveness. One important aspect of a measure's validity is responsiveness, the ability to detect a meaningful change in a clinical state.[1] Experts have published recommendations for pain outcome measures to be used in clinical trials, but few published studies have compared responsiveness across different pain outcomes measures.[2,3] Additional data on comparative responsiveness of pain measures are needed to inform choices about outcome measures for observational and experimental chronic pain research.

The Stepped Care for Affective Disorders and Musculoskeletal Pain (SCAMP) study included longitudinal assessment of pain with several measures of pain intensity and pain-related function. Our objective was to compare responsiveness of pain measures—including pain intensity, functional interference, and composite measures—among patients with moderate-severe persistent musculoskeletal pain enrolled in the SCAMP randomized clinical trial and parallel observational cohort study.

Methods

We used data from SCAMP, which enrolled 500 primary care patients with persistent (≥3 months) back, hip, or knee pain of at least moderate severity (Brief Pain Inventory severity ≥5). Participants were recruited from internal medicine clinics affiliated with an academic public hospital (n = 300) or Veterans Affairs (VA) hospital (n = 200) in Indianapolis. SCAMP enrolled patients in 2 concurrent studies, a randomized trial of combined depression medication and pain self-management versus usual care (n = 250) and an observational cohort study (n = 250). Inclusion criteria for the trial required PHQ-9 (Patient Health Questionnaire)scores of ≥10 (consistent with depression of at least moderate severity), whereas criteria for the cohort study required PHQ-9 scores of <8 (consistent with absence of clinical depression); eligibility criteria for the 2 studies were otherwise identical. Details of SCAMP design and enrollment[4] and primary outcomes of the trial[5] have been published. For these analyses, we included participants with data available at both baseline and 12-month follow-up (n = 427). We examined responsiveness for randomized trial and observational cohort participants separately to allow for comparison of findings in the 2 study populations, which differed in baseline characteristics and symptom trajectories. The SCAMP protocol was approved by the Indiana University institutional review board and Roudebush VA Medical Center Research and Development review committee.

Measures Evaluated

We assessed responsiveness of the following pain outcome measures:

- The Brief Pain Inventory (BPI) has been validated as a measure of chronic pain in primary care and other clinical populations.[6–9] The BPI includes 4 items that assess the intensity of pain currently and at its least, worst, and average during the past

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

week (rated from 0, “no pain,” to 10, “pain as bad as you can imagine”). An additional 7 items assess pain-related functional interference (rated from 0, “does not interfere,” to 10, “interferes completely”). We assessed responsiveness of the 4-item BPI severity and 7-item BPI interference scales, as well as that of the 2 following variations: (1) a BPI total score that includes all eleven items, and (2) a 3-item abbreviated scale that includes 1 intensity item and 2 functional items (PEG; average Pain intensity during the past week, pain interference with Enjoyment of life, and pain interference with General activity).[10] For each scale, the total score is the average of all items (range, 0–10; higher scores are worse).

- The Chronic Pain Grade Questionnaire (CPG) has been validated as a measure of pain severity in primary care, chronic pain, and general populations.[11–14] The CPG includes 6 items that rate intensity of current pain, average pain, and worst pain, as well as pain interference with activities (all rated from 0 to 10). Items other than current pain refer to the past 3 months. We assessed responsiveness of the CPG pain intensity scale (the first 3 items) and the CPG disability scale (the 3 interference items). For each scale, the total score is the average of the individual item scores multiplied by 10 (range, 0–100; higher scores are worse).
- The Roland Disability Questionnaire is a pain-specific measure of physical disability validated in patients with back and other chronic pain conditions.[15,16] It includes a checklist of 24 statements about current effects of pain on function. The total score is the number of items endorsed (range, 0–24; higher scores are worse).
- The Short Form (SF) Bodily Pain Scale is a validated subscale of the Medical Outcomes Study SF-36 questionnaire.[17,18] It includes 2 items, one that assesses intensity of pain (scale range: 1 “none” to 6 “very severe”) and one that assesses how much pain has interfered with work (scale range: 1 “not at all” to 5 “extremely”) over the past 4 weeks. Responses are transformed into a 0 to 100 score (lower scores are worse).

Reference Standard Measure

We used a patient-reported retrospective global rating of change as the reference standard for change in pain.[19] At 12 months, participants were asked, “Overall, since starting the study, would you say your pain is worse, about the same, or better?” Those who reported their pain was better were asked a second question, “How much better is your pain?” with the following response options: a little, somewhat, moderately, a lot, or completely better.

Other Measures

Medical comorbidity was assessed using a checklist of common medical conditions.[20] The Hopkins Symptom Checklist (SCL)-20 was used to assess depression symptom severity.[21,22]

Missing items in BPI or Roland measures were prorated if 1 to 2 values were missing. Missing items in SF bodily pain or CPG scales were prorated if 1 value was missing. If more values were missing, the participant was excluded from analysis of that scale. At baseline, the number of participants with at least one missing value was 6 (1.4%) for BPI, 5 (1.2%) for CPG, 9 (2.1%) for Roland and 0 for SF bodily pain. At 12 months, the number of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

participants with at least one missing value was 13 (3.0%) for BPI, 27 (6.3%) for CPG, 17 (4.0%) for Roland and 8 (1.9%) for SF bodily pain.

Analysis

Experts do not agree on a single preferred approach to responsiveness assessment, but recommend combining several approaches, including both anchor-based and distribution-based methods.[1,3,23–25] Retrospective patient-reported global ratings of change are the most common anchors used for responsiveness analysis in pain research. We used the 12 month global rating of change as the reference standard for calculating standardized response means (SRM) and receiver operating characteristic (ROC) curve analyses of responsiveness. We used a distribution-based method, the standard error of measurement (SEM), to estimate minimal clinically important change for each pain measure. Finally, to assess responsiveness of each measure to group differences in the randomized trial, we evaluated standardized effect size (SES) statistics stratified by trial arm (intervention and control).

Standardized Response Means

The SRM is an effect size index that is advocated for comparing responsiveness of measures because it includes the response variance in the denominator.[26] Participants were grouped into 3 categories according to their global rating of change at 12 months (worse, same, or better). For each measure, we calculated the SRM (12 month mean change score/SD of change) stratified by category of change.[27] We calculated 95% confidence intervals for the SRM with a bootstrapping procedure.[28] SRM values are unitless and therefore directly comparable between measures. They can be interpreted with Cohen's guidelines for two-group effect sizes (ie, 0.2 is small, 0.5 is moderate, and 0.8 is large).[29,30]

Area Under the ROC Curve

Deyo and Centor proposed that 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.[31] ROC curves plot sensitivity on the y-axis against (1 - specificity) on the x-axis for a measure compared with a reference standard. We calculated the area under the ROC curve (AUC) for each outcome measure using patient-reported improvement on the global rating of change as the anchor. We also calculated AUCs for a moderate improvement threshold (“moderately,” “a lot,” or “completely better”).[3,32] AUC values are interpreted as the probability of a measure correctly discriminating between patients who have improved and those who have not; the possible range is 0.5 (the same as chance) to 1.0 (perfect discrimination).

Standard Error of Measurement

Wyrwich and others developed evidence for the SEM as a distribution-based estimate of the minimal clinically important difference.[24,33,34] We calculated the SEM for each measure by multiplying the baseline standard deviation by the square root of 1 minus the baseline internal consistency coefficient (Cronbach's alpha). A change equal to one-SEM has been proposed as an estimate of minimal clinically important change, so we categorized

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

participants as follows: those with score improvement ≥ 1 SEM from baseline were categorized as better, those with score worsening ≥ 1 SEM as worse, and those with <1 SEM change as the same. We then calculated weighted kappa statistics[35] to explore the agreement between classification as better, the same, or worse by one-SEM criteria compared with classification by global rating of change criteria.

Effect Sizes Within the Trial

The ability to detect change associated with an efficacious intervention is a necessary characteristic of any measure used in clinical trials.[3] Primary results of the SCAMP randomized trial showed the intervention was effective in reducing pain and pain-related functional limitations.[5] To assess measures' responsiveness to the intervention, we evaluated effect sizes according to trial arm assignment (intervention or control). We calculated change scores (12 months minus baseline) and standardized effect sizes (SES = [intervention group change minus control group change]/SD of pooled change scores).[36] To assess whether measures performed similarly among participants with low back pain and those with knee or hip pain, we also stratified this analysis by pain location.

Statistical analyses were performed using SAS version 9.1 (SAS Institute Inc., Cary, NC).

Results

Baseline characteristics of participants in the randomized trial (those with pain and depression, n = 205) and observational cohort (those with pain and no depression, n = 222) are presented in Table 1. Due to successful randomization, characteristics of participants in the intervention and control arms of the trial were equivalent.[5] Compared with participants in the cohort, those in the trial were younger (mean age, 55.7 vs. 62.2, $P < 0.001$) and less likely to be retired (41.9% vs. 22.9%, $P < 0.001$). By design, the trial group had more depressive symptoms (mean SCL-20 score of 1.9, consistent with moderately severe depression) than did the cohort group (mean SCL-20 score 0.7, consistent with mild or no depression). On every pain measure, the trial group had worse mean baseline scores than the cohort group; this is expected because the presence of comorbid depression is consistently associated with worse pain severity.[37]

According to the pain global rating of change, 103 (24.1%) participants were better at 12 months, 90 (21.1%) were worse, and 234 (54.8%) were about the same. Rates of improvement were higher in the trial group (60, 29.3%) than in the cohort group (43, 19.4%). Rates of worsening were also slightly higher in the trial group (n = 49, 23.9%) than in the cohort group (n = 41, 18.5%).

Standardized Response Means

Table 2 shows SRM values for patients classified as worse, the same, and better at 12 months according to their global rating of change. Overall, the various measures had similar effect sizes within each category of change. Participants with improved pain according to the global rating had moderate-to-large improvement in scores, regardless of study group. SRM values among those who reported improvement were somewhat lower for the CPG and SF bodily pain than for other scales.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Among observational cohort participants, mean SRM values significantly differed between the worse and same groups and between the better and same groups for each measure, indicating that measures were able to distinguish between participants who did and did not report global change in pain, regardless of the direction of change. Among randomized trial participants, SRMs significantly differed between the better and same groups for each measure; however, values for those who reported feeling worse did not differ from those of participants who reported feeling "about the same" at the end of the trial. Cohort participants who reported feeling worse at 12 months had small to moderate worsening in scores; in contrast, trial participants who felt worse at 12 months had minimal or no change in scores on most measures. We secondarily conducted analyses using 6 month change data; results were similar to those at 12 months (not shown).

Area Under The ROC Curve

Table 3 shows the area under the ROC curve (AUC) for each measure, first using any improvement on the global rating as the reference standard and second, using moderate improvement on the global rating. Overall, AUC values were similar between measures, although SF bodily pain and CPG disability had the lowest values. Results within the cohort and trial groups were similar and most AUC values were consistent with fair to good discriminatory ability (range, 0.65– 0.85). The responsiveness of each measure to the presence of any improvement was about as good as its responsiveness to moderate improvement.

Standard Error of Measurement

Calculated SEM values did not differ substantially by study group (Table 4). We used the SEM, an estimate of minimal clinically important change, to categorize patients as better, the same, or worse for each measure. Agreement between classification defined by the one-SEM criteria and by the global rating of change anchor was generally fair.[35] For all measures in both study groups, one-SEM criteria classified fewer participants as the same and more as better than the global rating did. Both one-SEM and global rating criteria classified more trial participants than cohort participants as better. However, one-SEM criteria classified fewer trial participants than cohort participants as worse; whereas global rating criteria classified more trial than cohort participants as worse.

Effect Sizes Within the Randomized Trial

Measures detected a small-to-moderate intervention group effect, with SES ranging from 0.41 for the CPG intensity scale to 0.67 for the Roland disability questionnaire (Table 5). The CPG intensity and disability scales had the lowest overall responsiveness to intervention effect (overall SES 0.41 and 0.43, respectively). Among participants with knee or hip pain, responsiveness of SF bodily pain (SES, 0.39), and CPG intensity (SES, 0.32) was relatively poor. For the other measures, responsiveness to the intervention was comparable among participants with back pain and with knee/hip pain.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Discussion

Using data from the SCAMP randomized trial and observational cohort studies, we found that multiple pain outcome measures were responsive to change among primary care patients with persistent back and lower extremity joint pain. We found similar responsiveness, overall, for measures of pain intensity, pain-related function, and composite pain and function. One brief composite measure (the PEG) was similarly responsive to longer pain intensity and function measures; another (the SF bodily pain scale) had relatively poor responsiveness.

Keller et al previously found comparable responsiveness of BPI severity, BPI interference, CPG intensity, CPG disability, and SF bodily pain scales in an observational study of primary care patients with back pain or arthritis.[8] We found that measures of pain intensity and measures of pain-related function had similar responsiveness compared with patient-reported global improvement. However, BPI and CPG intensity measures appeared to be more responsive than their corresponding function measures to global worsening. Intensity and function measures were similarly responsive to the SCAMP intervention, with small-to-moderate effect sizes.

Guidelines for assessment of pain recommend measuring both pain intensity and pain-related function.[2] A drawback of using separate scales to measure each domain is the need to interpret 2 different numbers; this may complicate decision-making if the intensity and functional ratings are discordant. Using composite pain measures rather than separate intensity and function scales may simplify pain assessment and data analysis. Our finding that PEG and BPI composite measures integrating both domains into a single number performed comparably to separate BPI intensity and functional interference measures support the validity of this strategy.

Given sufficient validity, brief measures are desirable for their enhanced feasibility in both large studies and clinical settings. Our findings demonstrate tradeoffs between brevity and validity for the shortest scale we assessed, ie, the 2-item SF bodily pain scale. This scale had the lowest internal consistency and was less responsive than other measures according to several analyses. In contrast, responsiveness of the 3-item composite PEG scale was similar to that of the longer scales, suggesting that it may represent a reasonable compromise between feasibility and validity.

Recall window is another factor that may affect responsiveness. The 2 scales with the longest recall were the least responsive to intervention effect; CPG scales (3 month recall) had the lowest overall responsiveness and SF bodily pain (4 week recall) was the next least responsive. Responsiveness was better for measures that assessed current (Roland) or past week (BPI and PEG) symptoms.

Except for the SF bodily pain and CPG intensity scales, responsiveness to intervention effect did not differ substantially according to pain location (back versus hip/knee). This was true for the Roland questionnaire, even though it was originally developed for back pain assessment.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We found only fair agreement between the one-SEM-based classification and the retrospective anchor-based global rating classification of change. For all measures in both study groups, the one-SEM criteria classified more participants as improved and fewer as unchanged than the global change rating did. This may have occurred because our global change anchor was insensitive to truly minimal changes and picked up more clinically relevant moderate changes. In the trial group in particular, participants who reported their overall pain was “about the same” actually had small score improvements. Sensitivity of global change measures is affected by both measure (eg, response options) and patient (eg, pain chronicity) factors[38]; agreement with one-SEM criteria is likely to be affected by these factors as well.

Most of our findings were replicated in the 2 distinct SCAMP study groups, although the trial and cohort groups differed in ways that could potentially influence results. First, trial participants had moderate-severe depression at baseline and those in the cohort study did not. Comorbid depression is known to amplify the symptom intensity and disability associated with chronic pain[37] and may affect performance of pain measures. Second, half of the trial participants were exposed to an intervention, which may alter expectations in a way that particularly influences retrospective assessment of pain change, our reference standard. These factors likely affected apparent responsiveness to worsening in the trial group.

This study has several limitations. First, we lack a true external criterion standard for pain change. The reference standards we used, global rating of change classification and receipt of an efficacious trial intervention, are both imperfect. For example, global rating of change can be affected by “present state bias,” in which the rating of change is overly influenced by pain severity at the time of the rating.[3] Additionally, retrospective rating questions have not been standardized, although differences in question formatting may not substantially affect responsiveness results.[38] Our second reference standard, receipt of an efficacious trial intervention, also lacks true independence from the measures being evaluated because patient-reported outcomes were necessary to determine the intervention's efficacy. Ultimately, no patient-reported measure can be considered a “gold standard” or truly independent of the others. Nonetheless, patient-reported measures are the only valid tools we have to assess the inherently subjective phenomenon of pain. To increase confidence in our findings, we used combined approaches and looked for consistency in relative responsiveness of measures.

A second limitation is the use of multiple comparisons between multiple measures. Findings of differences between measures should be interpreted with caution and confirmed in future research. Another limitation is a likely ceiling effect for worsening of pain in our study population. Participants in SCAMP, especially those in the trial, had moderate-severe chronic pain at baseline and limited room for pain to worsen. Overall, change scores among those who reported improvement were larger in magnitude than change scores among those who reported worsening. Finally, these results are most relevant to the samples we studied. They may not be generalizable to other patient populations or pain conditions; in particular, findings should not be applied to patients who have communication or cognitive impairments or those in nursing home or hospital settings. However, because back pain and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

joint pain are the 2 most common types of chronic pain and because depression comorbidity is prevalent, the primary care populations we studied are particularly important for future pain research.

What conclusions can be drawn from our findings? First, all measures discriminated between study participants with and without improvement over 12 months, so choice among these measures for observational research may be based on criteria other than their relative responsiveness. Second, if a brief composite measure is desired to enhance feasibility, the 3-item PEG may be a more responsive option than the 2-item SF bodily pain scale. Third, although all measures detected the trial intervention effect, the CPG and SF bodily pain scales were less responsive and therefore may be less desirable as primary clinical trial outcome measures. Finally, our findings of modest agreement between one-SEM and global rating classifications of change support prior recommendations that researchers should consider both anchor and distribution-based methods of determining clinically important change in pain outcome measures.

Acknowledgments

Supported by National Institute of Mental Health grant R01 MH-071268 (to K.K.) and by HSR&D Research Career Development Awards from the Department of Veterans Affairs (to E.E.K. and M.J.B.).

References

1. Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Med Care. 2000; 38(suppl 9):II84–II90. [PubMed: 10982093]
2. Dworkin RH, Turk DC, Farrar JT, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain. 2005; 113:9–19. [PubMed: 15621359]
3. Dworkin RH, Turk DC, Wyrwich KW, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain. 2008; 9:105–121. [PubMed: 18055266]
4. Kroenke K, Bair M, Damush T, et al. Stepped Care for Affective Disorders and Musculoskeletal Pain (SCAMP) study: design and practical implications of an intervention for comorbid pain and depression. Gen Hosp Psychiatry. 2007; 29:506–517. [PubMed: 18022044]
5. Kroenke K, Bair MJ, Damush TM, et al. Optimized antidepressant therapy and pain self-management in primary care patients with depression and musculoskeletal pain: a randomized controlled trial. JAMA. 2009; 301:2099–2110. [PubMed: 19470987]
6. Cleeland, CS. Pain assessment in cancer. In: Osoba, D., editor. Effect of Cancer on Quality of Life. Boca Raton, FL: CRC Press; 1991. p. 293-305.
7. Cleeland CS, Nakamura Y, Mendoza TR, et al. Dimensions of the impact of cancer pain in a four country sample: new information from multidimensional scaling. Pain. 1996; 67:267–273. [PubMed: 8951920]
8. Keller S, Bann CM, Dodd SL, et al. Validity of the brief pain inventory for use in documenting the outcomes of patients with noncancer pain. Clin J Pain. 2004; 20:309–318. [PubMed: 15322437]
9. Tan G, Jensen MP, Thornby JI, et al. Validation of the brief pain inventory for chronic nonmalignant pain. J Pain. 2004; 5:133–137. [PubMed: 15042521]
10. Krebs EE, Lorenz KA, Bair MJ, et al. Development and initial validation of the PEG, a three-item scale assessing pain intensity and interference. J Gen Intern Med. 2009; 24:733–738. [PubMed: 19418100]
11. Elliott AM, Smith BH, Smith WC, et al. Changes in chronic pain severity over time: the Chronic Pain Grade as a valid measure. Pain. 2000; 88:303–308. [PubMed: 11068118]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

12. Smith BH, Penny KI, Purves AM, et al. The Chronic Pain Grade questionnaire: validation and reliability in postal research. Pain. 1997; 71:141–147. [PubMed: 9211475]
13. Von Korff M, Deyo RA, Cherkin D, et al. Back pain in primary care. Outcomes at 1 year. Spine. 1993; 18:855–862. [PubMed: 8316884]
14. Von KM, Ormel J, Keefe FJ, et al. Grading the severity of chronic pain. Pain. 1992; 50:133–149. [PubMed: 1408309]
15. Jensen MP, Strom SE, Turner JA, et al. Validity of the Sickness Impact Profile Roland scale as a measure of dysfunction in chronic pain patients. Pain. 1992; 50:157–162. [PubMed: 1408311]
16. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine. 1983; 8:141–144. [PubMed: 6222486]
17. McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): part II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care. 1993; 31:247–263. [PubMed: 8450681]
18. Ware JE, Gandek B. The SF-36 Health Survey: development and use in mental health research and the IQOLA Project. Int J Ment Health. 1994; 23:49–73.
19. Fischer D, Stewart AL, Bloch DA, et al. Capturing the patient's view of change as a clinical outcome measure. JAMA. 1999; 282:1157–1162. [PubMed: 10501119]
20. Perkins AJ, Kroenke K, Unutzer J, et al. Common comorbidity scales were similar in their ability to predict health care costs and mortality. J Clin Epidemiol. 2004; 57:1040–1048. [PubMed: 15528055]
21. Katon W, Von KM, Lin E, et al. Collaborative management to achieve treatment guidelines: impact on depression in primary care. JAMA. 1995; 273:1026–1031. [PubMed: 7897786]
22. Williams JW Jr, Stellato CP, Cornell J, et al. The 13 and 20 item Hopkins Symptom Checklist Depression Scale: psychometric properties in primary care patients with minor depression or dysthymia. Int J Psychiatry Med. 2004; 34:37–50. [PubMed: 15242140]
23. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol. 1997; 50:79–93. [PubMed: 9048693]
24. Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008; 61:102–109. [PubMed: 18177782]
25. Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol. 1997; 50:239–246. [PubMed: 9120522]
26. Katz JN, Larson MG, Phillips CB, et al. Comparative measurement sensitivity of short and longer health status instruments. Med Care. 1992; 30:917–925. [PubMed: 1405797]
27. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care. 1990; 28:632–642. [PubMed: 2366602]
28. Lowe B, Unutzer J, Callahan CM, et al. Monitoring depression treatment outcomes with the Patient Health Questionnaire-9. Med Care. 2004; 42:1194–1201. [PubMed: 15550799]
29. Cohen, J. Statistical Power Analysis for the Behavioral Sciences. Hills-dale, NJ: Lawrence Erlbaum Associates, Inc; 1988.
30. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care. 1989; 27(suppl 3):S178–S189. [PubMed: 2646488]
31. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chron Dis. 1986; 39:897–906. [PubMed: 2947907]
32. Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. Mayo Clin Proc. 2002; 77:371–383. [PubMed: 11936935]
33. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol. 1999; 52:861–873. [PubMed: 10529027]
34. Wyrwich KW, Nienaber NA, Tierney WM, et al. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care. 1999; 37:469–478. [PubMed: 10335749]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174. [PubMed: 843571]

36. Buchbinder R, Bombardier C, Yeung M, et al. Which outcome measures should be used in rheumatoid arthritis clinical trials? Clinical and quality-of-life measures' responsiveness to treatment in a randomized controlled trial. Arthritis Rheum. 1995; 38:1568–1580. [PubMed: 7488277]

37. Bair MJ, Robinson RL, Katon W, et al. Depression and pain comorbid-ity: a literature review. Arch Intern Med. 2003; 163:2433–2445. [PubMed: 14609780]

38. Lauridsen HH, Hartvigsen J, Korsholm L, et al. Choice of external criteria in back pain research: Does it matter? Recommendations based on analysis of responsiveness. Pain. 2007; 131:112–120. [PubMed: 17276006]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Baseline Characteristics of Study Participants by Study Group Enrollment

Characteristic	All Participants (n = 427)	Randomized Trial* (n = 205)	Observational Cohort* (n = 222)	P†
Age, mean yr (SD)	59.1 (13.0)	55.7 (11.4)	62.2 (13.7)	<0.001
Women, n (%)	228 (53.4)	110 (53.7)	118 (53.2)	0.917
Race, n (%)				
White	249 (58.3)	122 (59.5)	127 (57.2)	0.890
Black	163 (38.2)	76 (37.1)	87 (39.2)	
Other	15 (3.5)	7 (3.4)	8 (3.6)	
Employment status, n (%)				<0.001
Employed	99 (23.2)	50 (24.4)	49 (22.1)	
Unemployed or unable to work	188 (44.0)	108 (52.7)	80 (36.0)	
Retired	140 (32.8)	47 (22.9)	93 (41.9)	
Mean (SD) no. medical diseases	2.7 (1.4)	2.8 (1.5)	2.6 (1.4)	0.188
Pain location, n (%)				0.135
Back	235 (55.3)	121 (59.0)	114 (51.8)	
Hip or knee	190 (44.7)	84 (41.0)	106 (48.2)	
SCL-20 depression, mean (SD)	1.25 (0.84)	1.88 (0.65)	0.66 (0.50)	<0.001
BPI severity, mean (SD)	5.7 (1.8)	6.0 (1.8)	5.3 (1.8)	<0.001
BPI interference, mean (SD)	5.8 (2.4)	6.9 (2.1)	4.7 (2.2)	<0.001
BPI total, mean (SD)	5.7 (2.0)	6.6 (1.8)	4.9 (1.8)	<0.001
PEG, mean (SD)	6.0 (2.2)	6.9 (2.0)	5.2 (2.1)	<0.001
CPG intensity, mean (SD)	68.3 (16.7)	71.8 (17.0)	65.0 (15.9)	<0.001
CPG disability, mean (SD)	56.8 (29.4)	69.1 (25.2)	45.3 (28.5)	<0.001
Roland disability, mean (SD)	14.8 (5.5)	17.3 (4.3)	12.4 (5.4)	<0.001
SF bodily pain, mean (SD)	35.3 (18.1)	27.0 (15.3)	42.9 (17.0)	<0.001

*Randomized trial participants had moderate-severe pain and moderate-severe depression; observational cohort participants had moderate-severe pain and no clinical depression.

†*P* values are from χ^2 or *t* tests for comparisons between the randomized trial and observational cohort study groups.

BPI indicates Brief Pain Inventory; CPG, Chronic Pain Grade; SD, standard deviation; SCL-20, Hopkins Symptom Checklist-20; SF, Medical Outcomes Study Short Form-36.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Standardized Response Means According to Global Change Category at 12 Month for Trial, Observational, and Full Study Groups

	12 mo SRM* (95% CI)		
Pain Measure Global Change Category	**Randomized Trial (n = 205)†**	**Observational Cohort (n = 222)**	**Full Study (n = 427)**
BPI severity			
Worse	0.29 (0.00, 0.58)	0.75 (0.43, 1.06)	0.49 (0.28, 0.70)
Same	−0.02 (−0.23, 0.18)	0.08 (−0.08, 0.25)	0.04 (−0.08, 0.17)
Better	−0.99 (−1.25, −0.73)	−1.07 (−1.38, −0.76)	−1.00 (−1.20, −0.81)
BPI interference			
Worse	0.06 (−0.22, 0.35)	0.43 (0.11, 0.75)	0.24 (0.03, 0.45)
Same	−0.50 (−0.70, −0.30)	−0.09 (−0.26, 0.08)	−0.24 (−0.37, −0.11)
Better	−1.06 (−1.32, −0.79)	−0.69 (−1.00, −0.38)	−0.86 (−1.10, −0.66)
BPI total			
Worse	0.15 (−0.13, 0.44)	0.63 (0.31, 0.94)	0.37 (0.16, 0.58)
Same	−0.42 (−0.62, −0.22)	−0.04 (−0.21, 0.12)	−0.18 (−0.31, −0.05)
Better	−1.15 (−1.41, −0.90)	−0.99 (−1.30, −0.68)	−1.02 (−1.21, −0.82)
PEG			
Worse	−0.05 (−0.33, 0.24)	0.35 (0.04, 0.67)	0.15 (−0.06, 0.36)
Same	−0.49 (−0.69, −0.28)	−0.13 (−0.30, 0.04)	−0.25 (−0.38, −0.12)
Better	−1.14 (−1.41, −0.88)	−0.83 (−1.14, −0.52)	−0.99 (−1.19, −0.79)
CPG intensity			
Worse	0.56 (0.28, 0.85)	0.60 (0.28, 0.91)	0.57 (0.37, 0.78)
Same	−0.03 (−0.24, 0.17)	0.07 (−0.10, 0.24)	0.02 (−0.10, 0.15)
Better	−0.73 (−0.99, −0.46)	−0.68 (−1.00, −0.36)	−0.67 (−0.86, −0.48)
CPG disability			
Worse	0.14 (−0.14, 0.43)	0.37 (0.07, 0.70)	0.27 (0.06, 0.48)
Same	−0.25 (−0.45, −0.05)	−0.03 (−0.20, 0.14)	−0.12 (−0.25, 0.01)
Better	−0.94 (−1.20, −0.67)	−0.57 (−0.89, −0.25)	−0.77 (−0.97, −0.57)
Roland disability			
Worse	0.35 (0.06, 0.63)	0.57 (0.25, 0.89)	0.45 (0.24, 0.66)
Same	−0.29 (−0.50, −0.09)	−0.03 (−0.20, 0.14)	−0.11 (−0.24, 0.02)
Better	−1.09 (−1.35, −0.83)	−0.67 (−0.98, −0.36)	−0.90 (−1.09, −0.70)
SF bodily pain			
Worse	−0.17 (−0.45, 0.12)	−0.58 (−0.89, −0.27)	−0.36 (−0.57, −0.15)
Same	0.31 (0.11, 0.52)	0.17 (0.00, 0.34)	0.22 (0.09, 0.35)
Better	0.76 (0.50, 1.02)	0.67 (0.36, 0.98)	0.72 (0.52, 0.92)

*SRM = 12 month-baseline change score/SD of change score.

†In the trial group, 49 (23.9%) participants were categorized as worse, 96 (46.8%) as the same, and 60 (29.3%) as better. In the observational group, 41 (18.5%) were categorized as worse, 138 (62.2%) as the same, and 43 (19.4%) as better.

BPI indicates Brief Pain Inventory; CPG, Chronic Pain Grade; SD, standard deviation; SF, Medical Outcomes Study Short Form-36; SRM, standardized response mean.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

Accuracy of Pain Measures for Detecting Improvement in Randomized Trial and Observational Cohort Groups

	Accuracy for Detecting any Improvement		Accuracy for Detecting Moderate Improvement*	
	Randomized Trial AUC (SE)	Observational Cohort AUC (SE)	Randomized Trial AUC (SE)	Observational Cohort AUC (SE)
BPI severity	0.81 (0.036)	0.83 (0.032)	0.85 (0.035)	0.81 (0.038)
BPI interference	0.78 (0.040)	0.70 (0.043)	0.77 (0.046)	0.67 (0.052)
BPI total	0.81 (0.038)	0.78 (0.036)	0.81 (0.042)	0.76 (0.044)
PEG	0.78 (0.038)	0.73 (0.040)	0.79 (0.042)	0.70 (0.050)
CPG intensity	0.78 (0.039)	0.75 (0.043)	0.82 (0.040)	0.73 (0.055)
CPG disability	0.75 (0.040)	0.65 (0.044)	0.76 (0.043)	0.66 (0.053)
Roland disability	0.81 (0.037)	0.70 (0.044)	0.85 (0.038)	0.70 (0.050)
SF bodily pain	0.72 (0.044)	0.68 (0.046)	0.77 (0.043)	0.70 (0.054)

AUC is probability of correctly discriminating between patients who have improved and those who have not.

*Moderate improvement = global rating of “moderately,” “a lot,” or “completely” better.

AUC indicates area under the receiver operating characteristic curve; BPI, brief pain inventory; CPG, chronic pain grade; SF, Medical Outcomes Study Short Form-36; SE, standard error.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Change Classification by Global Rating and One-SEM Criteria for Randomized Trial and Observational Cohort Groups

			Change Classified by Global Rating*			
			Better, %	Same, %	Worse, %	
Randomized Trial (n = 205)			29.3	46.8	23.9	
Observational Cohort (n = 222)			19.4	62.2	18.5	
			Change Classified by One-SEM Criteria§			
Pain Measure	**Alpha†**	**SEM‡**	**Better, %**	**Same, %**	**Worse, %**	**Kappa¶**
BPI severity						
Trial	0.84	0.7	43.4	27.3	29.3	0.32
Cohort	0.81	0.8	29.3	38.3	32.4	0.31
BPI interference						
Trial	0.89	0.7	56.9	25.3	17.8	0.24
Cohort	0.85	0.8	37.9	31.5	30.6	0.20
BPI total						
Trial	0.90	0.6	52.7	25.9	21.5	0.29
Cohort	0.86	0.7	34.7	35.1	30.2	0.34
PEG						
Trial	0.76	1.0	44.3	41.9	13.8	0.33
Cohort	0.69	1.2	29.4	45.7	24.9	0.23
CPG intensity						
Trial	0.72	9.0	36.6	38.1	25.3	0.35
Cohort	0.62	9.8	27.5	43.6	28.9	0.27
CPG disability						
Trial	0.88	8.7	44.8	32.8	22.4	0.27
Cohort	0.87	10.3	32.1	41.3	26.6	0.14
Roland disability						
Trial	0.83	1.8	44.4	32.2	23.4	0.36
Cohort	0.87	1.9	37.4	29.3	33.3	0.18

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

			Change Classified by Global Rating*			
			Better, %	Same, %	Worse, %	
SF bodily pain						
Trial	0.59	9.8	46.3	33.0	20.7	0.19
Cohort	0.57	11.1	23.1	60.6	16.3	0.27

*Change classified by global rating at 12 month, in response to the following question: "Overall, since starting the study, would you say your pain is worse, about the same, or better?"

†Cronbach's alpha (a) coefficient at baseline.

‡SEM = SD × $\sqrt{(1 - a)}$ where SD indicates standard deviation at baseline.

§Change classified by one-SEM criteria as follows: better = score improved ≥1 SEM from baseline, same = score change <1 SEM from baseline, and worse = score worsened ≥1 SEM.

¶Weighted kappa statistic for agreement between one-SEM and global rating classification.

BPI indicates brief pain inventory; CPG, chronic pain grade; SF, Medical Outcomes Study Short Form-36; SEM, standard error of measurement.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 5

Effect Sizes of the Intervention Among Randomized Trial Participants Overall and According to Pain Site

	Overall (n = 205)			Back Pain (n = 121)			Knee or Hip Pain (n = 84)		
	Intervention Change (SD)*	Control Change (SD)*	SES†	Intervention Change (SD)	Control Change (SD)	SES	Intervention Change (SD)	Control Change (SD)	SES
BPI severity	1.2 (2.3)	−0.0 (1.8)	0.56	1.2 (2.3)	0.07 (1.7)	0.54	1.2 (2.4)	−0.2 (2.1)	0.58
BPI interference	2.1 (2.9)	0.6 (2.0)	0.59	2.1 (3.1)	0.7 (1.8)	0.54	2.2 (2.7)	0.4 (2.4)	0.66
BPI total	1.8 (2.5)	0.4 (1.8)	0.64	1.8 (2.6)	0.5 (1.5)	0.60	1.8 (2.3)	0.2 (2.1)	0.69
PEG	2.0 (2.7)	0.6 (2.0)	0.58	2.0 (2.8)	0.7 (1.7)	0.54	2.0 (2.5)	0.4 (2.3)	0.65
CPG intensity	7.9 (22.4)	−0.5 (17.6)	0.41	9.1 (21.2)	0.0 (13.9)	0.49	6.1 (24.4)	−1.2 (21.8)	0.32
CPG disability	17.1 (31.3)	4.6 (24.2)	0.43	17.6 (29.9)	6.4 (20.7)	0.43	16.3 (33.7)	2.2 (28.5)	0.44
Roland disability	3.5 (5.4)	0.3 (3.4)	0.67	3.6 (5.3)	0.3 (3.2)	0.71	3.3 (5.6)	0.3 (3.6)	0.60
SF bodily pain	−12.1 (21.5)	−2.1 (17.0)	0.50	−13.2 (22.8)	−1.3 (17.1)	0.57	−10.5 (19.6)	−3.3 (17.0)	0.39

*Change = baseline score – 12 month score.

†SES = (Intervention group change – control group change)/SD of pooled change score.

BPI indicates brief pain inventory; CPG, chronic pain grade; SD, standard deviation; SF, Medical Outcomes Study Short Form-36; SES, standardized effect size.