Clinical outcome assessment in clinical trials of chronic pain treatments

Kushang V. Patel,*, Dagmar Amtmann, Mark P. Jensen, Shannon M. Smith, Christin Veasley, Dennis C. Turk

Abstract
Clinical outcome assessments (COAs) measure outcomes that are meaningful to patients in clinical trials and are critical for determining whether a treatment is effective. The objectives of this study are to (1) describe the different types of COAs and provide an overview of key considerations for evaluating COAs, (2) review COAs and other outcome measures for chronic pain treatments that are recommended by the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) or other expert groups, and (3) review advances in understanding pain-related COAs that are relevant to clinical trials. The authors reviewed relevant articles, chapters, and guidance documents from the European Medicines Agency and U.S. Food and Drug Administration. Since the original core set of outcome measures were recommended by IMMPACT 14 years ago, several new advancements and publications relevant to the measurement or interpretation of COAs for chronic pain trials have emerged, presenting new research opportunities. Despite progress in the quality of measurement of several outcome domains for clinical trials of chronic pain, there remain some measurement challenges that require further methodological investigation.

Keywords: Clinical trials, Chronic pain, Outcome assessment, Patient-reported outcomes

1. Introduction
Clinical outcome assessments (COAs) measure outcomes that are meaningful to patients in clinical trials and are critical for determining whether a treatment is effective. Unlike a biomarker (e.g., hemoglobin A1c), COAs measure how a patient feels or the impact of a health condition on how the patient functions in daily life. Understanding the benefits and risks of a treatment and how regulatory agencies decide on labeling claims depends, in part, on the characteristics of the COA used in clinical trials. For chronic pain treatments, the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) has recommended 6 core outcome domains that should be considered when designing clinical trials: (1) pain, (2) physical functioning, (3) emotional functioning, (4) participant ratings of improvement and satisfaction with treatment, (5) symptoms and adverse events (AEs), and (6) participant disposition. The systematic collection and reporting of outcomes in these domains are valuable for adjudicating the efficacy of treatments and comparing results across trials, although not all of these domains are always measured with COAs per se. The objectives of this review article are to (1) describe the different types of COAs and provide an overview of key considerations for evaluating COAs, (2) review COAs and other outcome measures for chronic pain treatments that are recommended by IMMPACT or other expert groups for phase 2 and 3 clinical trials, and (3) review advances in understanding pain-related COAs that are relevant to clinical trials. In addition to relevant book chapters and peer-reviewed publications identified through PubMed and Google Scholar searches, guidance documents from the European Medicines Agency and the U.S. Food and Drug Administration (FDA) were reviewed because they aim to integrate best practices for patient-centered outcome assessment. Furthermore, the bibliographies of key publications were examined along with citation searches to identify additional studies. Note that the target patient populations for the selected set of outcome measures reviewed herein are adults with chronic pain conditions; however, the article by Palermo et al. included in this supplement provides a comprehensive review of trial design considerations for pediatric patient populations. Outcome assessment in adults with cognitive disorders, such as developmental disabilities or dementia, will not be reviewed, although pain assessment in older adults with dementia has been reviewed previously. The reader is referred to that review for information about the COA tools currently available to assess pain in this population.
2. Types of clinical outcome assessments

There are 4 major types of COAs: patient-reported outcomes (PROs), clinician-reported outcomes (ClinROs), observer-reported outcomes (ObsROs), and performance-based outcomes (PerfOs). Table 1 presents the FDA’s definition for each COA type. The first 3 types differ according to who is rating the concept that is being measured. A PRO records a research participant’s response to questions about their own health or functioning that is independent of a clinician’s or anyone else’s judgement. This definition is similar to the European Medicines Agency description of PROs. By contrast, ClinROs rely on the judgment of a trained professional about the health or functioning of a patient based on observation or examination. Observer-reported outcomes do not involve direct reporting from either patients or clinicians, but rather are observations of patients made by a person who generally does not have specialized clinical training (eg, family member or caregiver). This person can, however, receive training on what to observe. In contrast to the COAs that rely on a rater’s judgement, PerfOs measure a patient’s performance of a discrete task following an established protocol (eg, 6-minute walk test). Although PerfOs usually involve trained professional staff who systematically record a patient’s or study participant’s performance, they do not incorporate the assessor’s judgement or interpretation of performance. Importantly, outcome measures that are assessed with mobile health technologies (eg, wearable physical activity or sleep monitors) do not necessarily fit within the 4 major COA types as currently defined by the FDA, although they do provide a unique source of information that may be important in the evaluation of treatment effects (see section 4.2.4); whether such outcome approaches are ultimately classified as a subtype of ObsRO or will be classified as a distinct COA type remains to be seen.

3. Considerations for evaluating clinical outcome assessments

The selection of COAs as primary or secondary endpoints should be based on the anticipated effects of the intervention (eg, drug or nonpharmacological treatment) on specific aspects of how a patient feels or functions in daily life. The primary endpoint is used to determine treatment efficacy and/or safety, whereas the purpose of secondary endpoints is to enhance the interpretation of the primary endpoint result and/or to provide evidence for a specific mechanism underlying the treatment effect. Therefore, the measurement properties of COAs play a critical role in not only detecting a treatment effect but also in establishing treatment benefit for improving meaningful aspects of the study participant’s health. The key measurement properties for evaluating COAs include content validity, reliability, construct validity, responsiveness, and interpretability. A comprehensive discussion of these properties is beyond the scope of this article, but each are reviewed briefly below and have been discussed at length in numerous publications.

3.1. Content validity

Content validity is the degree to which the COA measures the concept of interest (ie, the domain or factor that the COA is purported to measure). Evidence for content validity is based primarily on both expert opinion and qualitative research engaging patients from the target population. A critical component of content validity is ensuring that the COA comprehensively covers the entire range of health issues (eg, symptoms or functional limitations) experienced by the patient population that is relevant to the concept of interest. Generally, content validity should be established before evaluating other measurement properties. Expert methodologists have published detailed guides on establishing the content validity of existing or newly developed PROs that can be adapted for other COA types.

3.2. Reliability

Reliability is the extent to which the COA is free of measurement error. In clinical trials, the reproducibility of a treatment effect depends, in part, on the reliability of the COA. Test–retest reliability is critical for PROs and PerfOs, whereas intrarater reliability and interrater reliability are particularly relevant for ClinROs and ObsROs. For test–retest reliability, the period between assessments should be carefully considered in the context of the population being studied and measurement construct—a prolonged time interval will increase the probability that a study participant might experience a real change in their health status or functioning, whereas too short an interval can result in participants simply recalling their initial (baseline) responses. For COAs with an interval scale of measurement, the intraclass correlation coefficient (ICC) is a statistic used to measure the degree of agreement between either repeated administrations (test–retest reliability) or multiple observers/raters (interrater reliability) of an outcome measure in the same set of
patients. The ICC can also be used to estimate intrarater reliability. By contrast, the kappa statistic is commonly used to measure agreement (test-retest, interrater, and intrarater reliability) for COAs with a nominal measurement scale. In addition to the reproducibility aspect of reliability, the internal consistency of PROs with multiple questionnaire items is also an important consideration because it reflects the interrelatedness (correlations) among the items, providing an indication of the extent to which the PRO items measure a single, latent (unobservable) construct.\textsuperscript{148} The Cronbach alpha coefficient is the most widely used statistic to assess internal consistency of multi-item PROs.\textsuperscript{27}

In addition to the traditional metrics based on classical test theory (CTT) that assess the reliability of the total score from a PRO or other type of COA, the precision of individual measurement items and the total score in distinguishing study participants across different levels of a latent construct is also important. Item response theory (IRT) is a family of statistical models that aim to explain the relationship between study participants’ responses to individual items on a measurement scale and the latent construct. The item information derived from IRT models can be summed to provide an overall assessment of the scale’s precision (for detailed discussion of IRT, see Ref. 114). Unlike the CTT-based reliability, the IRT-based information (the IRT equivalent of CTT reliability) is estimated at each point along the latent construct, recognizing that the amount of error is typically greater at the extremes of the continuum and providing important information for powering clinical trials.\textsuperscript{114}

3.3. Construct validity

Construct validity is the extent to which the COA demonstrates relationships with other measures, outcome domains, or patient characteristics that are consistent with a priori hypotheses.\textsuperscript{147} For example, to evaluate the convergent validity (a type of construct validity) of a new PRO of pain intensity among patients with chronic low back pain (CLBP), we would expect greater pain intensity would correlate positively with higher levels of emotional distress and pain-related interference with physical function. Discriminant (or divergent) validity is demonstrated when the COA is less strongly associated with measures of constructs that are theoretically more distal or unrelated. The extent to which a COA of interest correlates with a gold standard measure or legacy measure is criterion validity, a subtype of convergent validity.\textsuperscript{147} Known groups validity is another form of construct validity in which comparisons of a COA are made between different groups of patients that are known to differ in an expected way on the concept of interest. Generally, multiple studies are needed to cumulate evidence of construct validity for a COA, requiring time and resources for measurement development.

3.4. Responsiveness

Responsiveness is the ability to detect change over time in the concept of interest.\textsuperscript{147} For clinical trials to establish treatment benefit, it is critical for COAs to be able to detect improvement and worsening in the concept of interest across the entire measurement range for a given target patient population.\textsuperscript{147} Evidence for responsiveness of a COA can be generated from either clinical trials or observational studies by examining the relationships of change in COA scores with changes in other relevant outcome measures.\textsuperscript{116, 145} Sample size calculations for clinical trials depend, in part, on the responsiveness of the COA selected as the primary endpoint.\textsuperscript{116, 145}

3.5. Interpretability

In chronic pain treatment trials, PROs are the primary type of COA used as pain and other related symptoms are subjective experiences that are generally not known by anyone other than the patient or study participant. Unlike common clinical signs, such as systolic blood pressure, it is often difficult to meaningfully interpret change in PRO score values.\textsuperscript{25} In addition, it is widely recognized that a statistically significant between-group difference in PRO scores does not necessarily equate to a clinically meaningful difference.\textsuperscript{26, 39, 41} Furthermore, statistical significance does not indicate the extent to which individual patients experienced clinically meaningful improvement.\textsuperscript{25, 26, 36, 41} Therefore, to enrich the interpretation of PRO and other COA endpoint results, appropriate responder definitions should be established that identify clinically meaningful within-person change for the target patient population (it is important to note that clinically important within-person change does not equate to clinically important group difference because larger magnitudes of change are generally required for the former than the latter).\textsuperscript{38, 41}

Although there are a variety of anchor-based and distribution-based methods for identifying responder definitions, recent guidelines recommend triangulating from multiple methods and analyses (including use of multiple anchors) to derive either a single threshold value or a range of values for interpreting meaningful changes in COA scores.\textsuperscript{36} Based on several methodological studies, IMMPACT recommends reporting the proportion of patients with reductions in pain intensity of $\geq 30\%$ and $\geq 50\%$ in clinical trials of chronic pain treatment.\textsuperscript{41} Because responder analyses play an increasingly important role in regulatory decision-making, more methodological research is needed to define clinically meaningful change of COAs used in chronic pain trials.

4. Review of clinical outcome assessments for chronic pain

As noted earlier in section 1.0, IMMPACT recommends the assessment of 6 core outcome domains for clinical trials of chronic pain treatments.\textsuperscript{137} In the subsequent sections, we review the specific COAs recommended by IMMPACT\textsuperscript{37} and other expert groups, including the Outcome Measures in Rheumatology (OMERACT), Osteoarthritis Research Society International (OARSI), Neuropathic Pain Special Interest Group (NeuPSIG), National Institutes of Health (NIH) Task Force on Research Standards for Chronic Low Back Pain, and Veterans Health Administration Pain Measures Work Group. Relevant COAs that are increasingly used in the field will also be reviewed.

4.1. Core outcome domain of pain

Treatments for chronic pain can target and affect multiple aspects of the pain experience, including pain intensity, pain quality, and the use of rescue analgesic medications.\textsuperscript{36} Given that pain is experienced privately, PROs are generally used to assess pain intensity and quality. Reliable and valid pain assessment is critical for not only determining the effectiveness of chronic pain treatments, but also for diagnosing chronic pain conditions. Indeed, accurate pain assessment is a core component of dimension 1 of the Analgesic, Anesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks (ACTTION)-American Pain Society Pain Taxonomy (AAPT).\textsuperscript{48}

4.1.1. Pain intensity

Pain intensity is likely the most commonly assessed aspect of chronic pain in clinical practice and the most commonly
Numerous studies have demonstrated the reliability, validity, and electronic (with automated data capture) versions of the VAS. The distance in millimeters (mm) from the beginning of the line (“No pain” side) to the marking is the participant’s pain level. “The distance in millimeters (mm) from the beginning of the line (“No pain” side) to the marking is the participant’s pain level.” The distance in millimeters (mm) from the beginning of the line (“No pain” side) to the marking is the participant’s pain level. Each of these types of scales has strengths and weaknesses that should be considered when designing a chronic pain trial.74

4.1.1.3. Numerical rating scale for pain intensity

There are multiple versions of the VRS; but, in general, the VRS for pain intensity is a rank-ordered list of words describing different levels of pain that include extreme anchors and intermediate adjectives (eg, no pain, mild pain, moderate pain, and severe pain). Each descriptor is coded with a numerical value for data analysis (eg, 0, 1, 2, and 3 for no pain, mild pain, moderate pain, and severe pain, respectively). Respondents are asked to select the single descriptor that best represents their level of pain. Although the VRS is easy to administer and has demonstrated construct validity,68,74,123 the scoring method assumes that the VRS has an interval scale of measurement. However, it is unlikely that the magnitude of difference in pain intensity between “no pain” and “mild pain” is the same as the difference between “moderate” and “severe.” This can pose a challenge for interpreting change in VRS scores over time.

4.1.1.2. Visual analogue scale for pain intensity

In contrast to the VRS, the VAS has measurement scale properties approaching a ratio scale.111,113 The VAS asks respondents to mark their level of pain on a line (horizontal or vertical) that is usually 10 centimeters long with ends that are labeled with anchors such as “No pain” and “Worst imaginable pain.” The distance in millimeters (mm) from the beginning of the line (“No pain” side) to the marking is the participant’s pain level. There are paper-based, mechanical (with a sliding marker), and electronic (with automated data capture) versions of the VAS. Numerous studies have demonstrated the reliability, validity, and responsiveness of the VAS for pain intensity assessment in various chronic pain patient populations68,74,123, however, compliance with the VAS is lower than with VRS and NRS, particularly among older adults,68 potentially reflecting the greater motor, visual, and cognitive demands required for completing the VAS.

4.1.1.3. Numerical rating scale for pain intensity

The NRS is likely the most commonly used measure of pain intensity in clinical settings because it is easy to administer and score. Research participants are asked to rate their pain intensity on a 0 to 10 (or 20 or 100) scale where 0 usually represents “No pain” and 10 represents a descriptor indicating an extreme level of pain (eg, “Worst imaginable pain”). As with the VRS and VAS, there are several versions of the NRS. Some instruct the study participant to verbally report or write down the number that best reflects their pain intensity, whereas others present the numbers or boxes with numbers in ascending order and participants circle/ mark their pain level. The measurement properties of the NRS for pain intensity are robust and well established across multiple patient populations,68,74,123 although recent research suggests that the NRS may not have utility in populations of individuals with chronic pain from developing countries with low literacy rates.106 A clear strength of the NRS is its versatility with different modes of data collection, ranging from in-person interviews to telephone calls and interactive voice recording to mobile technology and wearable devices. However, in contrast to the VAS, it is unlikely that the 0 to 10 NRS has a ratio scale of measurement. For instance, a 3-point decline on the 0 to 10 NRS starting from a baseline of 8 probably does not have the same meaning as a 3-point decline starting at 4.

4.1.1.4. Picture or face scales for pain intensity

In addition to the NRS, VAS, and VRS, pain intensity can also be assessed by using drawings or photographs of different facial expressions of pain.74 Although the image or face scales were originally developed for use in adult populations with impairment in cognitive function or communication,124 they also have the potential to standardize pain assessment in the general adult population from across different countries and cultures. Indeed, in a recent study of adults with musculoskeletal pain in Nepal,106 the Faces Pain Scale-Revised (FPS-R)67 demonstrated adequate construct validity and was the most preferred scale followed by the VRS, VAS, and NRS. This study also highlighted the challenges of assessing pain in a non-Western country with a relatively low literacy rate and limited exposure to rating scales in daily life.106 For example, approximately a quarter of participants did not respond to the NRS or VAS for pain intensity despite receiving verbal instructions that were repeated up to 3 times, and a third responded to the NRS with a range of pain intensities rather than selecting a single value.106 In addition, 15% of participants volunteered different adjectives to describe their pain intensity rather than selecting one of the response options from the VRS (translated from English to Nepali).106 By contrast, the FPS-R had fewer problematic responses than the 3 other measures of pain intensity, suggesting that the FPS-R might be a more appropriate measure for use in multisite trials involving non-Western countries with low literacy. However, further research in such settings is required that evaluates different pain intensity assessments and participant training programs to enhance the reliability and validity of pain intensity scores.

4.1.1.5. Recommendations and advances for pain intensity assessment

Although the 4 types of pain intensity measures are highly correlated with one another,68 there are tradeoffs to consider when selecting one as a trial endpoint. Based on the findings of a literature review and a consensus meeting of experts, IMMPACT recommended the 0 to 10 NRS to assess pain intensity in chronic pain treatment trials.37 More recently, the Critical Path’s PRO Consortium completed a literature review and has also concluded that the 0 to 10 NRS is currently the most optimal pain intensity measure for clinical trials in adults without cognitive impairment.123 Despite these recommendations, there remain important challenges to standardized collection and reporting of pain intensity data in clinical trials. A systematic review by Smith et al.128 identified underreporting of critical elements of pain intensity assessment in clinical trials. For example, 20% of clinical trials published in leading pain journals between January 2011 and July 2012 had not clearly identified the pain intensity measure used, and 12% had not...
reported the anchors for the response options. Interestingly, another systematic review observed that among clinical research studies that used the 0 to 10 NRS, there were 14 different anchors used for the upper end of the scale (eg, “Worst pain imaginable,” “Unbearable pain,” and “Maximum pain”) and 11% of studies had not reported the response anchors. It is unclear to what extent pain intensity measures with different anchors have equivalent measurement properties (reliability, validity, and responsiveness). The review by Smith et al. also identified that only half of the trials reported the type of pain intensity rated (ie, average, least, worst, and current) and a third did not report the period rated (eg, past 12 or 24 hours). Based on these findings, Table 2 presents recommendations for standardizing the reporting of pain intensity assessments.

In efforts to improve the assay sensitivity of chronic pain trials (ie, the ability to detect a true treatment effect), investigators have sought to better understand and improve the measurement properties of pain intensity outcome measures. For example, in a secondary data analysis of a randomized controlled trial (RCT) of oxymorphone extended-release for CLBP, Jensen et al. sought to determine whether increasing the reliability of the pain intensity assessment (0–100 mm VAS of average pain in the last 24 hours collected in clinic) by increasing the number of ratings in composite outcome measures would increase assay sensitivity. Consistent with psychometric theory and previous research, the internal consistency of the composite measures increased as the number of ratings increased, although the incremental gain was small (the Cronbach alpha increased from 0.94 to 0.96 for composites with 2 and 9 ratings, respectively); however, there was no improvement in assay sensitivity (Cohen’s d = 0.57 for composites with 2 and 9 ratings). Indeed, a single pain rating had almost the same assay sensitivity (Cohen’s d = 0.52) as the composite measures of multiple ratings. This provocative finding suggested that the collection of multiple ratings may not be necessary, but the authors emphasized the need for replication and additional investigation. Accordingly, Stone et al. retrospectively analyzed data from a clinical trial of self-management for adults with osteoarthritis (OA) that collected 0 to 10 NRS average pain intensity data at home using interactive voice recording (recall period was the day of the call, “today”). Similar to the findings of Jensen et al., the internal consistency of pain intensity composites increased as the number of ratings increased, but the amount of gain was greater than that observed by Jensen et al., probably because the starting Cronbach alpha was much lower in the OA self-management trial than in the CLBP drug trial. However, unlike Jensen et al., improvements in pain intensity reliability were associated with increased assay sensitivity. Importantly, there are several methodological differences between these studies that can potentially explain differences in the study findings. Accordingly, additional studies are needed to further investigate the effect of increasing the internal consistency (reliability) of pain intensity assessment on the assay sensitivity of chronic pain trials.

More recent efforts by ACTTION to understand and ultimately improve assay sensitivity include the evaluation of composite responder outcomes of pain intensity and physical functioning in clinical trials of neuropathic pain. Patel et al. conducted an individual patient data analysis of 15 RCTs of duloxetine, gabapentin, and pregabalin for diabetic peripheral neuropathy (DPN) and postherpetic neuralgia (PHN). A composite responder outcome of ≥50% reduction in pain intensity, or a ≥20% reduction in pain intensity and ≥30% improvement in physical functioning was validated and cross-validated in different neuropathic pain conditions and treatments. Notably, this responder outcome had slightly lower number needed to treat than a standard responder outcome of ≥50% reduction in pain intensity, and was more favorably associated with patient global ratings of improvement. This initial study exploring composite responder outcomes of pain intensity and physical functioning in clinical trials of DPN and PHN requires replication in different contexts of use (eg, different patient populations and treatments).

In addition to harmonizing individual-level data from RCTs and conducting secondary analyses of outcome measures in chronic pain trials, ACTTION has published several systematic reviews and meta-analyses relevant to the design, analysis, and reporting of chronic pain treatment trials. Most recently, Smith et al. compared the assay sensitivity of change in average pain intensity vs change in worst pain intensity in RCTs of efficacious treatments for several chronic pain conditions. Previous research had suggested that peak (worst) pain and end (current) pain might influence overall (average) pain ratings, indicating that worst pain ratings might provide a more accurate assessment than average pain ratings. Furthermore, the FDA’s draft guidance on analgesic drug development recommends worst pain intensity ratings as a primary endpoint. However, the meta-regression analyses by Smith et al. did not show any consistent difference in assay sensitivity between average vs worst pain intensity ratings.

Finally, recognizing that patients often have idiosyncratic interpretations of PROs and are only given minimal instruction on how to respond or interpret PRO questionnaire items, ACTTION developed a program to train patients to be more accurate in their pain ratings. The program guided study participants to identify personal anchors for 0 to 10 NRS for pain intensity scale and educated participants on the meaning of “average” pain and its relationship to “least” and “worst” pain. Participants were also instructed to make their pain ratings in a quiet place, free of distractions. The training emphasized that participants should focus only on pain intensity when making their ratings and not on other symptoms (eg, fatigue or stress). In an

### Table 2

**Recommendations for reporting pain intensity assessments.**

| Recommendation                                                                                       |
|-------------------------------------------------------------------------------------------------------|
| Report the type of pain intensity assessment used (eg, VAS, NRS) with a description that clearly distinguishes the assessment from others (eg, for a VAS, describing the length of the line; for an NRS, reporting the range of possible ratings) |
| Report the definitions of anchors, except in cases where a well-known assessment is used verbatim and the anchors are easily referenced (eg, anchors for the 4 BPI pain intensity NRS items) |
| Report the frequency of administering the pain intensity assessment                                   |
| Report the period to be rated, except in cases where a well-known assessment is used verbatim and the period is easily referenced (eg, SF-MPQ PPI assesses present pain) |
| Report the type of pain intensity rated by participants (eg, average, usual, least, worst, current, and present) |
| Report the specific bodily area or pain condition to be rated; if none was specified, this should be stated |

BPI, Brief Pain Inventory; NRS, numerical rating scale; VAS, visual analogue scale.
In addition to the intensity or magnitude of pain, there are affective and sensory qualities of pain that can be modulated with therapeutics; therefore, it might be appropriate to measure pain affect or quality as an endpoint, depending on the intervention and patient population.

### 4.1.2. Pain quality

As described earlier, pain affect reflects the extent to which pain is disturbing, unpleasant, or causes distress.

Based on the correlation of pain affect with pain severity, it is important to measure pain affect or quality as an endpoint, depending on the intervention and patient population.

#### 4.1.2.1. Pain affect

The affect subscale of the McGill Pain Questionnaire is the most commonly used assessment with well-established measurement properties. For a historical perspective on over 40 years of research involving the MPQ and subsequent short forms, see the comprehensive review by Professor Chris Main. The original MPQ consisted of 20 classes of pain/symptom descriptors (2–6 descriptors in each class) that were used to identify different qualities of pain (eg, affective and sensory). The first short form-MPQ (SF-MPQ) improved the scaling and formatting of the assessment, reducing the number of pain descriptors to 15 that are each rated by patients or research participants using a VRS (response options: none, mild, moderate, and severe). The second short form (SF-MPQ-2) retained these original 15 descriptors, added 7 other descriptors to improve the assessment of neuropathic pain, and rescaled response options from the VRS to a 0 to 10 NRS (anchors: “none” and “worst possible”; symptom recall period: past week).

Exploratory and confirmatory factor analyses of the SF-MPQ and SF-MPQ-2 identify an affective scale with the following 4 response options from the VRS to a 0 to 10 NRS (anchors: “none” and “worst possible”; symptom recall period: past week).

- **SF-MPQ-2**
  - Pain intensity 0.7–0.9
  - (for a historical perspective on over 40 years of research involving the MPQ and subsequent short forms, see the comprehensive review by Professor Chris Main.)
  - The original MPQ consisted of 20 classes of pain/symptom descriptors (2–6 descriptors in each class) that were used to identify different qualities of pain (eg, affective and sensory).
  - The first short form-MPQ (SF-MPQ) improved the scaling and formatting of the assessment, reducing the number of pain descriptors to 15 that are each rated by patients or research participants using a VRS (response options: none, mild, moderate, and severe).
  - The second short form (SF-MPQ-2) retained these original 15 descriptors, added 7 other descriptors to improve the assessment of neuropathic pain, and rescaled response options from the VRS to a 0 to 10 NRS (anchors: “none” and “worst possible”; symptom recall period: past week).

#### 4.1.2.2. Sensory quality

The sensory qualities of pain can provide clues about the potential biopsychosocial mechanisms underlying the painful condition and help guide treatment decisions. There are several multi-item questionnaires with different strengths and weaknesses that can be used to assist with phenotyping research participants or, potentially, assess treatment efficacy.

The SF-MPQ-2 is designed to provide an overall assessment of pain-related symptoms by including a range of neuropathic and nonneuropathic pain descriptors. In addition to assessing pain affect, multiple studies on adults with different pain conditions have identified 3 other subscales: continuous pain (“throbbing pain,” “cramping pain,” “tingling pain,” “heavy pain,” and “tender”); intermittent pain (“shooting pain,” “stabbing pain,” “splitting pain,” “electric-shock pain,” and “piercing”); and neuropathic pain (“hot-burning pain,” “cold-freezing pain,” “pain caused by light touch,” “itching,” “tingling or pins and needles,” “numbness”). These subscales have demonstrated internal consistency, convergent and divergent validity, and responsiveness to treatment in a variety of pain conditions.

Researchers have also computed ICCs for test–retest reliability of the subscales that range from a low of 0.4 to 0.7 over a 3-month test interval in adults with knee OA to a high of 0.9 to 1.0 over a 3-day test interval in adults with chronic visceral pain. Although the factor structure of the SF-MPQ-2 is well established, the moderate-to-high correlations (r = 0.7–0.9) between the subscale scores raise some concern about whether they represent unique latent constructs.

In future research, it would be valuable to determine whether the subscales discriminate between different pain populations (eg, determine differences in the neuropathic pain subscale in adults with and without neuropathic pain conditions) or predict treatment response (ie, effect modification).

The Pain Quality Assessment Scale (PQAS) is the only other rigorously evaluated PRO that assesses a broad array of sensory qualities of pain. Originally based on the Neuropathic Pain Scale (NPS)—a brief (11-item), validated measure of neuropathic pain that discriminates between different pain conditions and detects treatment effects—the PQAS added 10 more items to improve the content validity for assessing neuropathic pain as well as nonneuropathic pain symptoms.

In total, there are 21 items in the PQAS that assess pain intensity (1 item), unpleasantness (1 item), sensory qualities (16 items), spatial characteristics (2 items rating intensity of deep and surface pain), and temporal pattern (1 item identifying intermittent, variable, or stable pain). With exception to the last categorical item measuring the temporal pattern of pain, all other items are assessed on a 0 to 10 NRS with a 1-week recall period and anchors of “No pain” and “The most intense pain sensation imaginable” (note: anchors vary depending on the pain descriptor/symptom being assessed). The wording of the PQAS (instructions and certain items) was revised slightly to improve understandability based on cognitive testing with chronic pain patients.

In adults diagnosed with carpal tunnel syndrome (CTS; n = 138), knee OA (n = 368), and LBP (n = 455), exploratory and confirmatory factor analyses of the spatial and sensory pain quality items identified 3 PQAS subscales: paroxysmal pain (shooting, sharp, electric, hot, and radiant), superficial pain (itchy, cold, numb, sensitive, and tingling), and deep pain (aching,
As described earlier, the NPS, PQAS, and SF-MPQ-2 are assessments consisting of neuropathic pain quality descriptors that have been used previously in neuropathic pain treatment trials. The Neuropathic Pain Symptom Inventory (NPSI) contains 10 symptom descriptors that are assessed on a 0 to 10 NRS with a 24-hour recall period and descriptor-specific anchors (eg, “No burning” to “Worst burning imaginable”). In addition, there are 2 temporal items that assess pain duration and the number of pain paroxysms in the past 24 hours. The factor structure of the NPSI has ranged from 3 to 5 dimensions, depending on the pain populations studied and analytic techniques, although evoked pain and deep pain were factors identified consistently across studies. Several studies have demonstrated that the NPSI has acceptable to excellent measurement properties. Furthermore, IMPACT has recommended the NPSI for phenotyping patients with neuropathic pain.

Several brief and easy-to-use screening tools for neuropathic pain are available, including the Leeds Assessment of Neuropathic Symptoms and Signs (LANSS), Douleur neuropathique en 4 questions (DN4), PainDETECT, and ID Pain. Similar to neuropathic pain assessments, these screening measures use symptom descriptors such as burning or electric shock to identify neuropathic pain. In general, these measures have adequate sensitivities and specificities, but some perform better depending on the characteristics of pain population and study goal. Importantly, a systematic review has identified limitations in the reliability and validity of these screening measures, particularly in their translations into different languages as well as their lack of diagnostic specificity. As noted by IMPACT, however, self-report screening tools should not replace a comprehensive clinical examination for diagnosing a neuropathic pain condition.

4.1.3. Rescue analgesics and concomitant pain treatments

Systematic monitoring of rescue medication use and concomitant use of pain treatments during the course of a chronic pain clinical trial is necessary for appropriately interpreting trial results and is recommended as a core outcome measure by IMPACT. As noted by Rowbotham and McDermott in their article in this supplement, pain relief derived from rescue medication can reduce assay sensitivity. Therefore, the amount (dose), date taken, and reason for the rescue medication use (eg, to differentiate relieving pain for the condition being investigated or some other reason such as a transient headache or dental pain) should be systematically recorded (ideally at the time the medication is taken). These data can then be analyzed as a secondary outcome (eg, time to initial rescue medication use or total amount taken through the course of the trial). Further research is needed to explore methods of integrating rescue medication use with other outcome measures such as pain intensity ratings.

4.2. Core outcome domain of physical functioning

For many adults, chronic pain negatively impacts their physical functioning and improvement in physical functioning is an important treatment goal. However, treatments designed to improve pain may not necessarily improve physical functioning. For example, in pooled analyses of 15 clinical trials involving first-line treatments for neuropathic pain, the correlation between heavy, dull, cramping, and throbbing). The internal consistencies of the subscales ranged from 0.69 to 0.85 within the different pain populations. Correlations between the subscales were not reported. Notably, study participants with CTS (ie, neuropathic pain) had statistically significant lower scores on the paroxysmal and deep pain subscales and higher surface pain scores than those with knee OA. The surface pain scores were also significantly higher in the CTS group compared with the LBP group. In terms of responsiveness, the individual items of PQAS were shown to improve with lidocaine patch 5% in patients with CTS. Among patients with moderate-to-severe neuropathic pain (peripheral neuropathy) who participated in an enriched enrollment randomized withdrawal trial of pregabalin, all 3 PQAS pain quality subscales improved from pretreatment to posttitration phases; however, pregabalin had a greater effect on the PQAS paroxysmal pain subscale than on the deep or surface pain scales, or on pain intensity. During the withdrawal phase, pregabalin (vs placebo) had the largest effect on paroxysmal pain followed by surface pain, but there were no statistically significant effects on deep pain or pain intensity. A similar pattern of results was observed with the PQAS subscales in an enriched enrollment randomized withdrawal trial of an extended-release formulation of oxymorphone for LBP, although there was a high discontinuation rate in the withdrawal phase of this trial. Taken together, these studies demonstrate that analgesics can have differential effects on pain quality as measured by PQAS. Furthermore, subsequent analysis of the pregabalin trial showed the paroxysmal and deep pain subscales at baseline were associated with treatment response but not with placebo response, illustrating the potential predictive value of phenotyping self-reported pain qualities.

Both the SF-MPQ-2 and PQAS provide a relatively brief (21–22 items), overall assessment of the sensory qualities of pain; however, each has unique strengths and weaknesses. A clear strength of the SF-MPQ-2 is that it reliably assesses pain affect (21–22 items), overall assessment of the sensory qualities of pain; however, each has unique strengths and weaknesses. A clear strength of the SF-MPQ-2 is that it reliably assesses pain affect (dose), date taken, and reason for the rescue medication use (ie, to differentiate relieving pain for the condition being investigated or some other reason such as a transient headache or dental pain) should be systematically recorded (ideally at the time the medication is taken). These data can then be analyzed as a secondary outcome (eg, time to initial rescue medication use or total amount taken through the course of the trial). Further research is needed to explore methods of integrating rescue medication use with other outcome measures such as pain intensity ratings.

4.1.2.2.1. Neuropathic pain assessment and screening

Comprehensive reviews on self-reported measures of neuropathic pain have been published previously. There are several assessment questionnaires specifically designed to monitor neuropathic pain symptoms as well as screening tools to assist clinicians identify neuropathic pain. Both assessment and screening questionnaires can also be used to phenotype potential treatment responders/nonresponders. We briefly review both types of self-reported measures for use in clinical trials of neuropathic pain. There are also ClinROs that can be used in combination with the self-reported measures to improve diagnostic accuracy of neuropathic pain.
change in pain intensity from baseline to posttreatment and change in physical functioning was small (−0.22 in DPN trials and −0.08 in PHN trials). For these reasons, IMMPACT recommends physical functioning as a core outcome domain.

In pain research, the terms “physical functioning” and “pain interference” are often used interchangeably; however, it is important to distinguish between them to appropriately interpret trial results. Physical functioning refers to one’s ability to perform activities that require physical action, such as self-care (eg, bathing, dressing, and eating), walking indoors or outdoors, or climbing stairs. Pain interference, however, is the extent to which pain impedes one’s ability to perform or participate in activities, ranging from basic physical activities to more complex social activities. Some pain interference measures cover a wider range of activity domains, whereas others specifically assess pain interference with physical functioning. Therefore, when considering outcome measures, it is important to not only differentiate measures of physical functioning from pain interference, but also general measures of pain interference from specific measures of pain interference with physical functioning.

4.2.1. Types of clinical outcome assessments for physical functioning

Multiple measurement approaches can be used to assess physical functioning. Patient-reported outcome measures assess the study participant’s perception of their own abilities to perform activities, whereas PerfOs objectively assess a participant’s physical capacity to perform a standardized, discrete task. Although PROs and PerfOs can provide complementary information on a research participant’s physical functioning, neither of these types of COAs objectively measure the participant’s physical functioning in their home and/or work environments. However, wearable devices with accelerometers can be used to collect the research participant’s daily activity pattern in their real-world environment. It is likely that each type of assessment provides a different perspective on the participant’s functioning, and triangulating from them can provide a more comprehensive assessment. In the sections below, we review examples of different types of COAs for physical functioning and pain interference with physical functioning.

4.2.2. Patient-reported outcomes of physical functioning and pain interference

There are several generic and pain condition-specific PROs of physical functioning and pain interference. Generic measures facilitate the comparison of results across different clinical trials and pain populations as well as facilitate cost-effectiveness evaluation, whereas condition-specific measures have greater content validity for the population of interest and therefore will be more responsive to treatment effects. For these reasons, IMMPACT recommends the collection of both generic and condition-specific measures, if available. Because several groups have published reviews or recommendations on PROs of physical functioning and pain interference, we will highlight recommended measures in the following 2 sections.

4.2.2.1. Generic measures of physical functioning and pain interference

The Short Form-36 (SF-36) is one of the most commonly used PROs of health-related quality of life that measures 8 domains, including physical functioning. The physical function subscale consists of 10 items that ask participants, “Does your health now limit you in these activities? If so, how much?” Activities range from bathing or dressing oneself to running. Each item has 3 response options: “Yes, limited a lot”; “Yes, limited a little”; and “No, not limited at all.” Scores from each item are summed and transformed to a 0 to 100 scale, where higher scores indicate better physical functioning. The measurement properties of the SF-36 physical functioning subscale have been examined in multiple pain populations and are generally acceptable to excellent.

Both the Brief Pain Inventory (BPI) and MPI have pain interference scales with well-established measurement properties and are recommended by IMMPACT as core outcome measures. The MPI has 9 pain interference items with a 0 to 6 NRS, whereas the BPI has 7 items with a 0 to 10 NRS. Interestingly, although the BPI pain interference scale is typically analyzed with a single score based on factor analyses supporting unidimensionality, more recent studies using Rasch analysis and confirmatory factor analysis suggest that there are 2 separate subscales: “physical interference” (items: general activity, walking ability, and normal work) and “affective interference” (items: mood, relationships with others, and enjoyment of life), leaving out the item on sleep. Further analysis of the BPI factor structure with clinical trial data are needed to determine whether only the 3 physical/activity-related items are sufficient for measuring pain interference with physical functioning.

Building off of the SF-36, BPI, MPI, and many other legacy measures, the NIH PROMIS has developed item banks using IRT to comprehensively measure a variety of health-related domains, including physical functioning and pain interference. The IRT-based item banks can be administered by computer adaptive tests that assess the full spectrum of physical functioning and pain interference, from low to high. Published or custom short forms can also be used to measure physical functioning and pain interference. The reliability and validity of PROMIS physical functioning and pain interference measures have been established in several pain populations, and emerging studies are demonstrating the responsiveness of the short-form versions of these measures. Furthermore, there is now a crosswalk table available to transform the BPI pain interference scores to PROMIS pain interference short-form scores (crosswalks for scores from other PROMIS and legacy measures are available at http://www.prosetastone.org/Pages/default.aspx.)

4.2.2.2. Condition-specific measures of physical functioning and pain interference

In addition to generic measures, there are several condition-specific pain interference PROs, primarily for musculoskeletal conditions. For example, a recent systematic review of PROs for physical functioning or pain interference in adults with nonspecific LBP had identified 17 measures. After applying CONSensus-based Standards for the selection of health Measurement Instruments (COSMIN) methodology and undergoing a 2-round Delphi survey process with experts, the Oswestry Disability Index version 2.1 and 24-item Roland-Morris Disability Questionnaire were recommended as a core outcome measure for clinical trials on LBP. Importantly, this systematic review also highlighted that the content validity of many commonly used PROs of physical functioning and pain interference in adults with LBP is inadequate, which can potentially decrease the responsiveness of an outcome measure and reduce assay sensitivity.
Similar findings have been reported for PROs of physical functioning and pain interference in hip or knee OA. For example, although the Western Ontario and McMaster Universities Arthritis Index (WOMAC) is a widely used and recommended PRO in adults with hip or knee OA, systematic reviews find that most studies evaluating WOMAC do not meet the COSMIN criteria for modern measurement standards. This may, in part, reflect poor quality of study reporting rather than poor measurement qualities of the PRO. Future research evaluating the measurement properties of a new or existing COA should attend to the COSMIN criteria. Finally, in contrast to the abundance of condition-specific measures of physical functioning or pain interference for chronic musculoskeletal pain, recent systematic reviews have reported a dearth of measures for neuropathic pain. Thus, there is a need to either modify existing PROs or develop new ones that assess functional limitations that are attributable to specific neuropathic pain conditions because generic measures (eg, SF-36 physical function subscale) may not capture them.

4.2.3. Performance-based outcomes

Both IMMPACT and OMERACT recognize that PerfOs can provide complementary information to PROs. An advantage of PerfOs is that they assess one’s physical capacity in a standardized manner, eliminating behavioral adaptations (eg, changes in the frequency or how a person performs a task) or environmental modifications that can influence a person’s report of their ability to perform activities. There are a number of PerfOs used in clinical pain research and trials. A systematic review applied COSMIN methodology to assess the measurement properties of 21 PerfOs of physical functioning in adults diagnosed with hip or knee OA. Based on the results of the systematic review and the consensus of expert clinicians and researchers, OARSI recommends the following PerfOs assessments for hip or knee OA: the 30-second chair-stand test, 40-m fast-paced walk test, a stair-climb test, timed up-and-go test, and 6-minute walk test. Of these, the first 3 were recommended as the minimal core set of PerfO assessments for hip or knee OA or after arthroplasty. Subsequent methodological investigation of these PerfOs generally supports the OARSI recommendations. It should be noted that the OARSI-recommended set of PerfO assessments and other PerfOs might be appropriate for other chronic pain conditions, but their measurement properties must be established in the populations of interest, as recommended by IMMPACT and OMERACT.

4.2.4. Accelerometer-based outcomes

As described earlier, there are strengths and weaknesses to PROs and PerfOs for assessing physical functioning. A shared weakness is that neither of these COAs provides an objective, time-stamped assessment of the daily activities a person actually does in their own real-world environment. With advances in mobile technology and greater acceptance of wearable devices, accelerometry can provide a unique window to understand how individuals with chronic pain function physically and potentially respond to treatments. Although a number of observational studies have examined accelerometer-measured physical activity in adults with chronic pain, relatively few clinical trials have collected accelerometer data to assess physical activity. One exception is a randomized, placebo-controlled, crossover trial of celecoxib for knee OA in which the accelerometer data were combined with pain intensity ratings as a composite outcome measure. There was a greater difference in responder rates between celecoxib and placebo using the pain-activity composite outcome than using pain ratings alone. Although this finding is exciting and provocative, there is a need to further develop accelerometry assessment protocols in clinical trials of chronic pain as well as develop and validate novel outcome measures that capitalize on the richness of the raw signal data.

4.3. Core outcome domain of emotional functioning

It is widely recognized that pain and psychological distress often co-occur. Indeed, the co-occurrence of pain and depression likely ranges between 30% and 50%. Furthermore, longitudinal studies have demonstrated bidirectional effects between pain and depression, and psychological distress or psychiatric comorbidity can reduce the effectiveness of pain treatments. Accordingly, it is important to measure emotional functioning in chronic pain trials as a potential endpoint or as a treatment effect modifier.

The Beck Depression Inventory (BDI) and the Profile of Mood States have been recommended as core outcome measures for chronic pain trials. Both measures have well-established reliability and validity and have been used extensively in clinical trials, but these 2 measures have some weaknesses, including the length of the Profile of Mood States (65 items) and the BDI does not measure anxiety. However, the Hospital Anxiety and Depression Scale (HADS) provides a more efficient alternative with a total of 14 items (7 for anxiety and 7 for depression). In addition, the HADS was designed for use in a general medical setting rather than for a psychiatric patient population, excluding somatic symptoms that can be confounded by illness and side effects of treatment. The measurement properties of the HADS were recently evaluated in a randomized, double-blind, parallel-group study of tapentadol immediate-release vs oxycodone immediate-release in 666 adults with acute low back pain. The HADS demonstrated good-to-excellent internal consistency, convergent validity, predictive validity, and responsiveness to treatment, and the data fit the prespecified factor structure. Accordingly, these results indicate that the HADS is a valid PRO assessment of anxiety and depression in adults with acute low back pain. Recently, it was recommended by IMMPACT as a core phenotyping measure to assess general negative affect. Future research should replicate these findings and extend them to other pain populations.

The Patient Health Questionnaire-9 (PHQ-9) is a 9-item PRO that is used to screen for depression, assess the severity of depression, and monitor depression treatment in clinical settings. Each item measures 1 of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition’s (DSM-IV’s) criteria for depressive disorder diagnoses. The psychometric properties of the PHQ-9 are well-established. The PHQ-9 has been used in several clinical trials, and the 2-item version (PHQ-2) is recommended by the Veterans Health Administration Pain Measures Work Group, which prioritized strong psychometric properties and pragmatic aspects (brevity) in recommending outcome measures.

The NIH PROMIS has developed item banks for depression and anxiety based on items from legacy measures, such as the BDI, HADS, and PHQ-9. These item banks can be administered using computerized adaptive testing, or fixed-length short forms are available. As with pain interference and physical functioning, the PROMIS depression and anxiety measures have demonstrated excellent measurement properties and
crosswalks with legacy measures have been established. The NIH Task Force on Research Standards for Chronic Low Back Pain recommends the 4-item PROMIS short form for depression.

4.4. Core outcome domain of participant global ratings improvement

Global ratings of improvement and treatment satisfaction are self-reported measures that provide research participants the opportunity to integrate multiple aspects of their treatment experience into a single assessment, weighing treatment benefits (eg, improvements in pain, physical functioning, and emotional functioning) and harms (ie, side effects and AEs). In addition, these ratings can be used to gauge what study participants consider a meaningful treatment response. Indeed, responses to the global improvement or treatment satisfaction assessment serve as anchors for establishing the clinically important difference in pain intensity and other outcomes. Notably, improvements in pain intensity and physical functioning are 2 consistent aspects of the treatment response that patients with OA, CLBP, and FM consider personally meaningful. There are several single-item ratings (Likert scale) used at the end of clinical trials to assess patient improvement or satisfaction with treatment, including the Patient Global Assessment of Treatment Satisfaction, the Patient-Rated Global Assessment of Response to Therapy, and the Patient Global Impression of Change. The Patient Global Impression of Change with 7 ordinal response options (ranging from “Very much worse” to “Very much improved”) is recommended by IMMPACT as a core outcome measure.

4.5. Core outcome domain of reporting adverse events

In any type of clinical trial, it is critically important to systematically assess, analyze, and report AEs. This information is essential to determine the research participant’s overall treatment experience during the course of the clinical trial as well as to inform clinical decision-making. Therefore, ACTTION has conducted a series of systematic reviews to make evidence-based recommendations on methods of assessing, analyzing, and reporting AEs in pain treatment trials. Despite the expansion in 2004 of the Consolidated Standards of Reporting Trials (CONSORT) statement with a set of standard harms-reporting recommendations, the reporting of AEs after 2005 has remained inadequate in pharmacologic, nonpharmacologic, and intravenous pain treatment trials. In an effort to improve AE reporting, Smith et al. developed a comprehensive checklist of reporting items that captures clinically relevant AE information that are necessary to draw conclusions about the participants’ AE experience as well as the trial’s overall AE assessment (see Table 6 in Ref. 130).

4.6. Core outcome domain of participant disposition

In accordance with IMMPACT recommendations and CONSORT guidelines, the disposition of study participants should be systematically recorded to facilitate the transparent reporting of trial results. This includes detailed information on participants’ progression throughout the trial, from screening for eligibility to withdrawal from follow-up status. The article by Gewandter et al. in this supplement provides a clear description of the reporting requirements of pain clinical trials as well as provides a pain-specific CONSORT checklist.

5. Development of new clinical outcome assessments

Under the Prescription Drug User Fee Act (PDUFA) VI, the FDA is preparing guidance on Patient-Focused Drug Development, which is defined as, “a systematic approach to help ensure that patients’ experiences, perspectives, needs, and priorities are captured and meaningfully incorporated into the development and evaluation of medical products throughout the medical product life cycle.” The central role of patients throughout the drug development process has implications for COA. It will be critical to engage patients and define the intended treatment benefit as a specific aspect of how a patient feels or functions. In the case that no appropriate COA sufficiently measures this meaningful health aspect, then the FDA has a COA Qualification Program for applicants to enter into a multi-stage process of developing and validating a new COA that measures a well-defined COI in a specific context of use (ie, the intended population, setting, and manner of use). If approved, then the “COA qualification represents a conclusion that within the stated context of use, results of the assessment can be relied upon to measure a specific concept and have a specific interpretation and application in drug development and regulatory decision-making.” Currently, ACTTION is pursuing the COA qualification process for a new PRO measure of pain intensity (QUALITE-Pain: QUALified for Therapeutic Evaluations of Pain) and a new accelerometer-based physical activity tool (PAACT: Physical Activity Accelerometry Assessment for Analgesic Clinical Trials).

6. Summary and recommendations for future research

In summary, COAs play an important role in the evaluation of treatment efficacy and safety. We have reviewed several outcome domains and measures for clinical trials of chronic pain treatments that were recommended by expert groups. This review also identified opportunities to improve the quality of outcome assessment. For example, greater efforts are needed to standardize the reporting of critical elements of pain intensity assessment (see Table 2 and Ref. 128) as well as AEs. In addition, there remain opportunities to further evaluate the effects of training research participants to improve the accuracy of pain ratings in clinical trials. Finally, although the outcome measures reviewed in this article generally have acceptable-to-excellent measurement properties based on CTT, several measures have not been evaluated using modern psychometric methods such as IRT. In future research, the COSMIN checklists and taxonomy of measurement properties should be considered when evaluating existing, modified, or newly developed COAs.

Disclosures

D. Amtmann and M.P. Jensen participated in the development of the Patient Reported Outcomes Measurement Information System (PROMIS) pain interference measure, but did not receive any royalty payments for that work. M.P. Jensen codeveloped the Pain Quality Assessment Scale (PQAS) that is reviewed in this article, and receives royalties from licensing fees paid by funded researchers who use the measure (clinicians and unfunded researchers can use the PQAS without paying a licensing fee). D.C. Turk codeveloped the Multidimensional Pain Inventory (MPI) and the second short-form McGill Pain Questionnaire (SF-MPQ-2), but does not receive any royalty payments for the use of these.
measures. The remaining authors have no conflict of interest to declare.

Article history:
Received 26 February 2019
Received in revised form 14 July 2019
Accepted 1 August 2019
Available online 21 January 2021

References
[1] Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, Cellai D, Rothrock N, Keefe F, Callahan L, Lai JS. Development of a PROMIS bank to measure pain. PAIN 2010;150:173–82.
[2] 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.
[3] Askew RL, Cook KF, Keefe FJ, Nowinski CJ, Cellai D, Revicki DA, DeWitt EM, Michaud K, Tremble DL. Amtmann D, A PROMIS measure of neuropathic pain quality. Value Health 2016;19:623–30.
[4] Askew RL, Cook KF, Revicki DA, Cellai D, Amtmann D. Evidence from diverse clinical populations supported clinical validity of PROMIS pain interference and pain behavior. J Clin Epidemiol 2016;73:103–11.
[5] Askew RL, Kim J, Chung H, Cook KF, Johnson KL, Amtmann D. Development of a crosswalk for pain interference measured by the BPI and PROMIS pain interference short form. Qual Life Res 2013;22:7269–76.
[6] Attal N, Bouhassira D, Baron R. Diagnosis and assessment of neuropathic pain through questionnaires. Lancet Neurol 2018;17:456–66.
[7] Attal N, Fermanian C, Fermanian J, Lantéri-Minet M, Alchaar H, Bouhassira D. Neuropathic pain: are there distinct subtypes depending on the aetiology or anatomical lesion? PAIN 2006;138:343–53.
[8] Bair MJ, Robinson RL, Katon W, Kroenke K. Depression and pain comorbidity: a literature review. Arch Intern Med 2005;165:2343–45.
[9] Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561–71.
[10] Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. The S-LANSS score for outcome measures for physical functioning in patients with low back pain. J Pain Res 2016;9:251–5.
[11] Bennett MI, Smith BH, Torrance N, Potter J. The LANSS Pain Scale: the Leeds assessment of neuromuscular and neuropsychiatric symptom scale. J Pain 2010;11:162–71.
[12] Bennett M. The LANSS Pain Scale: the Leeds assessment of neuromuscular and neuropsychiatric symptom scale. J Pain 2010;11:162–71.
[13] Bennett M. The S-LANSS score for identifying pain of predominantly neuropathic origin: validation for use in clinical and postal research. J Pain 2005;6:149–58.
[14] Bouhassira D, Attal N, Aichaar H, Bourefou B, Brochet B, Bruxelle J, Cunin G, Fermanian J, Giniès P, Grun-Overdyking A, Jafari-Schlupe H, Lantéri-Minet M, Laurent B, Mick G, Serrie A, Valade D, Vicaut E. Comparison of pain syndromes associated with nervous or somatic lesions and development of a new neuropathic pain diagnostic questionnaire (DPN4). PAIN 2005;115:1833–40.
[15] Bennett M. The LANSS Pain Scale: the Leeds assessment of neuromuscular and neuropsychiatric symptom scale. J Pain 2010;11:162–71.
[16] Bennett M, Smith BH, Torrance N, Potter J. The LANSS Pain Scale: the Leeds assessment of neuromuscular and neuropsychiatric symptom scale. J Pain 2010;11:162–71.
[17] Bennett M, Smith BH, Torrance N, Potter J. The LANSS Pain Scale: the Leeds assessment of neuromuscular and neuropsychiatric symptom scale. J Pain 2010;11:162–71.
[18] Bennett M, Smith BH, Torrance N, Potter J. The LANSS Pain Scale: the Leeds assessment of neuromuscular and neuropsychiatric symptom scale. J Pain 2010;11:162–71.
[19] Bennett M, Smith BH, Torrance N, Potter J. The LANSS Pain Scale: the Leeds assessment of neuromuscular and neuropsychiatric symptom scale. J Pain 2010;11:162–71.
Sampaio C. Interpreting the clinical importance of group differences in chronic pain clinical trials: IMPACT recommendations. PAIN 2009; 146:238–44.

Dworkin RH, Turk DC, Revicki DA, Harding G, Coyne KS, Peirce-Sands BH, Davenport E, Burke LB, Cowan P, Farrar JT, Hertz S, Max MB, Rappaport BA, Melzack R. Development and initial validation of a revised and expanded version of the Short-form McGill Pain Questionnaire (SF-MPQ-2). PAIN 2009; 144:35–42.

Dworkin RH, Turk DC, Trudeau JJ, Benson C, Biondi DM, Katz NP, Kim M. Validation of the Short-form McGill Pain Questionnaire (SF-MPQ-2) in acute low back pain. J Pain 2015; 16:357–66.

Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Clesen SD, Farrar JT, Haythornthwaite JA, Jensen MP, Kersting TD, Ader DN, Brandenburg N, Burke LB, Celli D, Chandler J, Cowan P, Dimitrova R, Dionne R, Hertz S, Jadad AR, Katz NP, Kehlet H, Kramer JD, Manning DC, McCormick C, McDermott MP, McQuay HJ, Patel S, Porter L, Quessy S, Rappaport BA, Rafuichkol C, Revicki DA, Rothman M, Schmader KE, Stacey BR, Staufler JW, von Stein T, White RE, Witter J, Zavisic S. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMPACT recommendations. J Pain 2009; 10:95–101.

Edwards RR, Cahalan C, Calahan C, Mensing G, Smith M, Haythornthwaite JA. Pain, catastrophizing, and depression in the rheumatic diseases. Nat Rev Rheumatol 2011; 7:216–24.

Edwards RR, Dworkin RH, Turk DC, Angst MS, Dionne R, Freeman R, Hamilton P, Karkkainen A, Arensd-Nielsen S, Attal N, Baron R, Brel J, Bujanover S, Burke LB, Carr D, Choppial SP, Cowan P, Eeghnor M, Fillingim RB, Gerwens JD, Katz NP, Kopecke EA, Markand DN, Nomikos G, Porter L, Rappaport BA, Rice AS, Scavone JM, Scholz J, Simon LS, Smith SM, Tobias J, Tockarshevsky T, Veasly C, Versavel M, Wasan AD, Wen W, Yamitzky D. Patient phenotyping in clinical trials of chronic pain treatments: IMPACT recommendations. PAIN 2016; 157:1851–71.

European Medicines Agency. Committee for medicinal products for human use (CHMP): reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products. European Medicines Agency: London, United Kingdom, 2005.

European Medicines Agency. Appendix 2 to the guideline on the evaluation of antineuralnancer medicinal products in man. The use of patient-reported outcome (PRO) measures in oncology studies. European Medicines Agency: London, United Kingdom, 2016.

Farbarn JC, Couper J, Davies JB, O’Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy 1980; 66:271–8.

Fernandez E, Turk DC. Nociception and emotion: separation and synthesis. Psychol Bull 1992; 112:205–17.

Fillingim RB, Loeser JD, Baron R, Edwards RR. Assessment of chronic pain: domains, methods, and mechanisms. J Pain 2016; 17:710–20.

Freynhagen R, Wirth A, Gockel U, Torg AR. painDETECT: a new screening questionnaire to identify neuropathic components in patients with back pain. Curr Med Res Opin 2006; 22:1911–20.

Fries JF, Celli D, Rose M, Krishnan E, Bruce B. Progress in assessing treatment of neuropathic pain. Clin J Pain 2012; 28:683–6.

Gageur JJ, Mullins M, Huang H, Mannac-Dabie D, Ghabbarian A, Elloff B, Mirza F, Aybana M. A systematic review of measurement properties of patient-reported outcome measures used in patients undergoing total knee arthroplasty. J Arthroplasty 2017; 32:1688–97.

Gagnier JJ, Mullins M, Huang H, Zeitouni-Nuzzo S, Attal N, Baron R, Brel J, Bujanover S, Burke LB, Celli D, Chandler J, Cowan P, Dimitrova R, Dionne R, Hertz S, Jadad AR, Katz NP, Kotter H, Kramer JD, Manning DC, McCormick C, McDermott MP, McQuay HJ, Patel S, Porter L, Quessy S, Rappaport BA, Rauschikol C, Revicki DA, Rothman M, Schmader KE, Stacey BR, Staufler JW, von Stein T, White RE, Vitter J, Zavisic S. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMPACT recommendations. J Pain 2009; 10:95–101.

Geisser ME, Clauw DJ, Strand V, Gendreau RM, Palmer R, Williams DA. Contributions of change in clinical status parameters to Patient Global Impression of Change (PGIC) scores among persons with fibromyalgia treated with milnacipran. PAIN 2010; 149:375–8.

Gewandter JS, Eisenach JC, Gross RA, Jensen MP, Kellie FJ, Lee DA, Turk DC. Checklist for the preparation and review of pain clinical trial publications: a pain-specific supplement to CONSORT. PAIN Rep 2019; 4:e621.

Gould EM, Jensen MP, Victor TW, Gamholtto AM, White RE, Galer BS. The pain quality response profile of oxymorphone extended release in the treatment of low back pain. Clin J Pain 2009; 25:116–22.

Gracely RH, McGrath P, Dubner R. Validity and sensitivity of ratio scales of sensory and affective verbal pain descriptors: manipulation of affect by diazepam. Pain 1978; 3:19–20.

Guy W. Public health service alcohol, drug abuse, and mental health administration. ECDEU assessment manual for psychopharmacology, revised, 1976. Rockville: US Department of Health, Education, and Welfare, 1976. pp. 217–22. DHEW Publication no. [ADM] 76–338.

Haanpäa M, Attal N, Backonja M, Baron R, Bennett M, Bouhassira D, Cruccu G, Hansson P, Haythornthwaite JA, Iannetti GD, Jensen TS, Kaupilla T, Nurmikko T, Rice AS, Rowbotham M, Serra J, Sommer C, Smith BH, Treede RD. NeuPSIG guidelines on neuropathic pain assessment. PAIN 2011; 152:14–27.

Hadjistavropoulos T, Herr K, Pirikach KM, Craig KD, Gibson SJ, Lukas A, Smith JH. Pain assessment in elderly adults with dementia. Lancet Neurol 2013; 12:431–40.

Harriss K, Dawson J, Gibbons E, Lim CR, Beard DJ, Fitzpatrick R, Price AJ. Systematic review of measurement properties of patient-reported outcome measures used in patients undergoing hip and knee arthroplasty. Patient Rep Outcome Meas 2016; 7:101.

Herrington KS, Cheng JC, Bosma RL, Rogachov A, Kim JA, Davis KD. Beyond negative pain-related psychological factors: resilience is related to lower pain affect in healthy adults. J Pain 2017; 18:1117–28.

Hicks CL, von Baeyer CL, Spafford P, van Kortiaar I, Goodenough B. The Faces Pain Scale—revised: toward a common metric in pediatric pain measurement. PAIN 2001;93:173–83.

Hjermstad MJ, Fayers PM, Haugen DF, Caraceni A, Hanks GW, Lode H, Fainsinger R, Aas N, Kaasa S; European Palliative Care Research Collaborative (EPCR): comparative study of numerical rating scales, verbal rating scales, and visual analogue scales for assessment of pain intensity in adults: a systematic literature review. J Pain Symptom Manage 2011;41:1073–93.

Hunsinger M, Smith SM, Rothstein D, McKeown A, Parkhurst M, Hertz S, Katz NP, Lin AH, McDermott MP, Rappaport BA, Turk DC, Dworkin RH. Adverse event reporting in nonpharmacologic, noninterventional pain clinical trials: ACTTION systematic review. PAIN 2014; 155:2253–62.

Jensen MP, Gamholtto AM, Bolongero JA, Alon A, Smugar GS, Galer BS, Hewett DJ. The pain quality response profile of pregabalin in the treatment of neuropathic pain. Clin J Pain 2012;28:683–6.

Jensen MP, Gamholtto AM, Olayeye DO, Oleka N, Nalamachu SR, Galer BS. The pain quality assessment scale assessment of pain quality in carpal tunnel syndrome. J Pain 2006; 7:823–32.

Jensen MP, Hu X, Potts SL, Gould EM. Single vs composite measures of pain intensity: relative sensitivity for detecting treatment effects. PAIN 2013;154:534–8.

Jensen MP, Johnson LE, Gertz KJ, Galer BS, Gamholtto AM. The words patients use to describe chronic pain: implications for measuring pain. PAIN 2013;154:2722–8.

Jensen MP, Karoly P. Self-report scales and procedures for assessing pain in adults. In: Turk DC, Melzack R, editors. Handbook of pain assessment. New York: Guilford Press, 2011.

Jensen MP, Lin CP, Kupfer AE, Galer BS, Gamholtto AM. Cognitive testing and revision of the pain quality assessment scale. Clin J Pain 2013;29:400–10.

Jensen MP, McFarland CA. Increasing the reliability and validity of pain intensity measurement in chronic pain patients. PAIN 1993;55:196–203.

Jensen MP, Wang W, Potts SL, Gould EM. The meaning of global outcome measures in pain clinical trials: more than just change in pain intensity. Clin J Pain 2013;29:899–905.

Kachooei AR, Ebrahimzadeh MH, Erfani-Sayyar R, Salehi M, Eriazi R, Razi S. Short-form McGill pain questionnaire-2 (SF-MPQ-2): a cross-cultural adaptation and validation study of the Persian version in patients with knee osteoarthritis. Arch Bone Jt Surg 2015;3:45–50.

Kahaner D, Fredrickson BL, Schreiber CA, Riedelmeier DA. When more pain is preferred to less: adding a better end. Psychol Sci 1993;4: 201–5.

Kerns RD, Turk DC, Rudy TE. The West haven-yale multidimensional pain inventory (WHYMPI). PAIN 1985;23:345–56.
[126] Smith SM, Amtmann D, Askew RL, Gewandter JS, Hunsinger M, Jensen MP, McDermott MP, Patel KV, Williams M, Bacci ED, Burke LB, Chambers CT, Cooper SA, Cowan P, Desjardins P, Etropolski M, Farrar JT, Gilron I, Huang IZ, Katz M, Kerns RD, Kopecky EA, Rappaport BA, Resnick M, Strand V, Vanhove GF, Veasley C, Versavili M, Wasan AD, Turk DC, Dworkin RH. Pain intensity rating training: results from an exploratory study of the ACTTION PROTECT system. PAIN 2016;157: 1056–64.

[127] Smith SM, Chang RD, Pereira A, Shah N, Gilron I, Katz NP, Lin AH, McDermott MP, Rappaport BA, Rowbotham MC, Sampaio C, Turk DC, Dworkin RH. Adherence to CONSORT harms-reporting recommendations in publications of recent analgesic clinical trials: an ACTTION systematic review. PAIN 2010;152:215–21.

[128] Smith SM, Hunsinger M, McKeown A, Parkhurst M, Allen R, Kopko S, Lu Y, Wilson HD, Burke LB, Desjardins P, McDermott MP, Rappaport BA, Turk DC, Dworkin RH. Quality of pain intensity assessment reporting: ACTTION systematic review and recommendations. J Pain 2015;16: 299–305.

[129] Smith SM, Jensen MP, He H, Kitt R, Koch J, Pan A, Burke LB, Farrar JT, McDermott MP, Turk DC, Dworkin RH. A comparison of the assay sensitivity of average and worst pain intensity in pharmacologic trials: an ACTTION systematic review and meta-analysis. J Pain 2018;19: 952–960.

[130] Smith SM, Wang AT, Katz NP, McDermott MP, Burke LB, Coplan P, Gilron I, Hertz L, Lin AH, Rappaport BA, Rowbotham MC, Sampaio C, Sweeney M, Turk DC, Dworkin RH. Adverse event assessment, analysis, and reporting in recent published analgesic clinical trials: ACTTION systematic review and recommendations. PAIN 2013;154: 997–1008.

[131] Smuck M, Kao MC, Brar N, Martinez-Ith A, Choi J, Tomkins-Lane CC. Does physical activity influence the relationship between low back pain and obesity? Spine J 2014;14: 209–16.

[132] Stone AA, Schneider S, Broderick JE, Schwartz JE. Single-day pain assessments as clinical outcomes: not so fast. Clin J Pain 2014;30: 739–43.

[133] Tan G, Jensen MP, Thomby JL, Shanti BF. Validation of the brief pain inventory in chronic nonmalignant pain. Pain 2004;5: 123–6.

[134] Tsurkov AM, Monstrey P, Patel KV, Turk DC, Dworkin RH, Beatard-D, Clauw DJ, Gignac MA, Markman JD, Williams DA, Bujanover S, Burke LB, Carr DB, Choy EH, Conaghan PG, Cowan P, Farrar JT, Freeman R, Gewandter J, Gilron I, Gol V, Gover TD, Haddox JD, Kerns RD, Kopecky EA, Lee DA, Malamut R, Mease P, Rappaport BA, Simon LS, Singh JA, Smith SM, Strand V, Tugwell P, Vanhove GF, Veasley C, Walco GA, Wasan AD, Witter J. Assessment of physical function and participation in chronic pain clinical trials: IMPACT/POMERACT recommendations. PAIN 2016;157:1836–50.

[135] Treister R, Laval OD, Shecter JD, Khurana N, Bothmer J, Harte MC, Amtmann D, Askew RL, Gewandter JS, Hunsinger M, Jensen MP, McDermott RH. A comparison of the assay sensitivity of average and worst pain intensity in pharmacologic trials: an ACTTION systematic review and meta-analysis. J Pain 2018;19: 952–960.

[136] Trudeau J, Van Inwegen A, Parkhurst M, Allen R, Kopko S, Lu Y, Wilson HD, Burke LB, Desjardins P, McDermott MP, Rappaport BA, Turk DC, Dworkin RH. Quality of pain intensity assessment reporting: ACTTION systematic review and recommendations. J Pain 2015;16: 299–305.

[137] Stone AA, Schneider S, Broderick JE, Schwartz JE. Single-day pain assessments as clinical outcomes: not so fast. Clin J Pain 2014;30: 739–43.

[138] Tan G, Jensen MP, Thomby JL, Shanti BF. Validation of the brief pain inventory in chronic nonmalignant pain. Pain 2004;5: 123–6.

[139] Tsurkov AM, Monstrey P, Patel KV, Turk DC, Dworkin RH, Beatard-D, Clauw DJ, Gignac MA, Markman JD, Williams DA, Bujanover S, Burke LB, Carr DB, Choy EH, Conaghan PG, Cowan P, Farrar JT, Freeman R, Gewandter J, Gilron I, Gol V, Gover TD, Haddox JD, Kerns RD, Kopecky EA, Lee DA, Malamut R, Mease P, Rappaport BA, Simon LS, Singh JA, Smith SM, Strand V, Tugwell P, Vanhove GF, Veasley C, Walco GA, Wasan AD, Witter J. Assessment of physical function and participation in chronic pain clinical trials: IMPACT/POMERACT recommendations. PAIN 2016;157:1836–50.

[131] Smuck M, Kao MC, Brar N, Martinez-Ith A, Choi J, Tomkins-Lane CC. Does physical activity influence the relationship between low back pain and obesity? Spine J 2014;14: 209–16.

[132] Stone AA, Schneider S, Broderick JE, Schwartz JE. Single-day pain assessments as clinical outcomes: not so fast. Clin J Pain 2014;30: 739–43.

[133] Tan G, Jensen MP, Thomby JL, Shanti BF. Validation of the brief pain inventory in chronic nonmalignant pain. Pain 2004;5: 123–6.

[134] Tsurkov AM, Monstrey P, Patel KV, Turk DC, Dworkin RH, Beatard-D, Clauw DJ, Gignac MA, Markman JD, Williams DA, Bujanover S, Burke LB, Carr DB, Choy EH, Conaghan PG, Cowan P, Farrar JT, Freeman R, Gewandter J, Gilron I, Gol V, Gover TD, Haddox JD, Kerns RD, Kopecky EA, Lee DA, Malamut R, Mease P, Rappaport BA, Simon LS, Singh JA, Smith SM, Strand V, Tugwell P, Vanhove GF, Veasley C, Walco GA, Wasan AD, Witter J. Assessment of physical function and participation in chronic pain clinical trials: IMPACT/POMERACT recommendations. PAIN 2016;157:1836–50.

[135] Treister R, Laval OD, Shecter JD, Khurana N, Bothmer J, Harte MC, Amtmann D, Askew RL, Gewandter JS, Hunsinger M, Jensen MP, McDermott RH. A comparison of the assay sensitivity of average and worst pain intensity in pharmacologic trials: an ACTTION systematic review and meta-analysis. J Pain 2018;19: 952–960.

[136] Trudeau J, Van Inwegen A, Parkhurst M, Allen R, Kopko S, Lu Y, Wilson HD, Burke LB, Desjardins P, McDermott MP, Rappaport BA, Turk DC, Dworkin RH. Quality of pain intensity assessment reporting: ACTTION systematic review and recommendations. J Pain 2015;16: 299–305.

[137] Stone AA, Schneider S, Broderick JE, Schwartz JE. Single-day pain assessments as clinical outcomes: not so fast. Clin J Pain 2014;30: 739–43.