A systematic review of the utility of unidimensional and functional pain assessment tools in adult postoperative patients

Reham M. Baamer[1,2], Ayesha Iqbal[1], Dileep N. Lobo[3,4], Roger D. Knaggs[1,5], Nicholas A. Levy[6], Li S. Toh[1]

1. Division of Pharmacy Practice and Policy, School of Pharmacy, University of Nottingham, Nottingham, UK

2. Pharmacy Practice Department, Faculty of Pharmacy, King Abdul-Aziz University, Jeddah, Saudi Arabia

3. Nottingham Digestive Diseases Centre, National Institute for Health Research, Nottingham Biomedical Research Centre, Nottingham University Hospitals NHS Trust and University of Nottingham, Queen's Medical Centre, Nottingham, UK

4. MRC Versus Arthritis Centre for Musculoskeletal Ageing Research, School of Life Sciences, University of Nottingham, Queen's Medical Centre, Nottingham, UK

5. Pain Centre Versus Arthritis, University of Nottingham, Nottingham, UK

6. Department of Anaesthesia and Perioperative Medicine, West Suffolk Hospital NHS Foundation Trust, Bury St. Edmunds, UK

Correspondence to:
Professor D. N. Lobo
Nottingham Digestive Diseases Centre
National Institute for Health Research (NIHR) Nottingham Biomedical Research Centre
Nottingham University Hospitals NHS Trust and University of Nottingham
Queen's Medical Centre
Nottingham NG7 2UH, UK
Tel: +44-115-8231149
Fax: +44-115-8231160
Email: Dileep.Lobo@nottingham.ac.uk

Short title: Postoperative pain scores

Word count (excluding abstract, references, tables, and figure legends): 4414

No. of Figures: 1

No. of Tables: 4

No. of Refs: 79

Supplementary document: 1

This paper was presented at the Annual Meeting of the Surgical Research Society in March 2021 and a preliminary version has been published in abstract form – *Br J Surg* 2021; 108 (S5): znab282.044b. doi 10.1093/bjs/znab282.044.

ABSTRACT

Background: In this systematic review we aimed to appraise the evidence relating to the measurement properties of unidimensional tools to quantify pain after surgery. Furthermore, we wished to identify tools used to assess interference of pain with functional recovery.

Methods: Four electronic sources (MEDLINE, EMBASE, CINAHL, PsycINFO) were searched in August 2020. Two reviewers independently screened articles and assessed risk of bias using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist.

Results: Thirty-one studies with a total of 12498 participants were included. Most of the studies failed to meet the methodological quality standards required by COSMIN. Studies of unidimensional assessment tools were underpinned by low quality evidence for reliability (5 studies), and responsiveness (7 studies). Convergent validity was the most studied property (13 studies) with moderate to high correlation ranging from 0.5 to 0.9 between unidimensional tools. Interpretability results were available only for the visual analogue scale (7 studies) and numerical rating scale (4 studies). Studies on functional assessment tools were scarce in which only one study included an 'Objective Pain Score', a tool assessing pain interference with respiratory function and had low-quality for convergent validity.

Conclusions: This systematic review challenges the validity and reliability of unidimensional tools in patients after surgery. We found no evidence that any one unidimensional tool has superior measurement properties in assessing postoperative pain. In addition, because promoting function is a crucial perioperative goal, psychometric validation studies of functional pain assessment tools are needed to improve pain assessment and management.

The protocol was registered (No. CRD42020213495) with the PROSPERO database and can be accessed at https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=213495.

Key words:

COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN); functional pain assessment tool; pain scores; postoperative pain; tool utility; unidimensional pain assessment

INTRODUCTION

Patients experience acute pain after surgery due to tissue damage and inflammation at the operation site.[1-3] Careful assessment of pain by a valid and reliable tool[4] is the first step towards a rational choice of analgesic therapy[5] which is essential for ensuring patient comfort, mobility, satisfaction and reducing healthcare costs.[6] Most commonly used tools for the assessment of postoperative pain are unidimensional and assess only pain intensity.[4] These include the visual analogue scale (VAS),[7] numerical rating scale (NRS),[8] verbal rating scale (VRS),[9] sometimes referred as verbal descriptor scale (VDS),[10] and faces pain scales (FPS).[11] They are quick to administer and do not encroach on the time required for usual care.[12]

Despite their extensive use, the reliance on these unidimensional tools as the sole approach to measuring pain is currently insufficient as the cut-off points commonly used by healthcare providers do not reflect the patient's desire for additional analgesics.[13, 14] Furthermore, patients have reported difficulties in describing the complexity of their pain experience by a single numerical value, descriptive words or as a mark on a line.[12] Striving to lower pain intensity scores to zero as suggested by the "Pain as the 5th Vital Sign" campaign has not improved pain outcomes,[15-17] and resulted in increased opioid analgesic use in the post-anaesthesia care unit.[17] Furthermore, Vila *et al*.[18] highlighted the potential hazard associated with a pain score-based treatment algorithm in increasing the prevalence of sedation-related side effects by more than twofold. Treating pain as the 5th vital sign has been abandoned now as it may have contributed to the current US opioid epidemic.[19, 20]

Restoration of function by allowing the patient to breathe, cough, ambulate and turn in bed is important for postoperative pain relief.[21, 22] Therefore, assessing the functional impact of pain, which includes patient-centred objective assessment by a healthcare provider who judges if the pain prevents the patient from performing activities that help accelerate recovery, could be an appropriate alternative to achieve better pain assessment.[23] Hence, options to treat pain will be used to maximize functional capacity, rather than striving to reduce the patient's postoperative pain score to below a specified numerical value.[4, 20]

Despite being used widely, the validity, reliability, and utility of unidimensional pain assessment tools for postoperative patients have not been reviewed systematically. The aim of this systematic review was to appraise the available evidence concerning the measurement properties of different unidimensional and functional pain assessment tools when used to assess postoperative pain in hospitalised adults.

METHODS

We performed this systematic review according to COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) (http://www.cosmin.nl/) guidelines, and reported it according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement guidelines.[24]

Search Strategy

We performed a systematic search of the MEDLINE, EMBASE, PsycINFO (all via OVID) and CINAHL (via EBSCOhost) databases from their inception to August 2020. Our search strategy consisted of four search concepts: 1) measurement properties or outcome terms, 2) pain assessment tool terms, 3) acute postoperative pain and 4) limits (English language or English translation, human adults ≥18 years old). We combined the first three using the Boolean operator AND, which works as a conjunction to narrow the search to include our specific three search concepts resulting in more focused results. This was then combined the result string with the fourth concept to limit the results. We performed these steps separately for each pain assessment tool. We carried out backward citation tracking as well by checking the reference lists from eligible studies. The comprehensive search strategy used is provided in **Appendix S1**.

Inclusion Criteria

We included any of the following pain measurement tools to assess acute pain in hospitalised adult patients from all surgical specialties: unidimensional pain assessment tools [including the numerical pain rating scale, verbal rating scale, visual analogue scale, faces scales (Wong-Baker FACES, Faces Pain Scale-Revised)], and functional pain assessment tools included any tool that helps assess acute pain based on its interference with functional activity, including walking, breathing, turning in bed and coughing. Included functional pain assessment tools could be used objectively by the clinician or when self-reported by patients.

We included instrument validation or instrument evaluation types of studies. Any studies that included at least one or more of the instruments to evaluate postoperative pain and assessed at least one of the nine measurement properties identified by COSMIN taxonomy: internal consistency, test-retest reliability, measurement error, content validity, structural validity, construct validity, hypothesis testing, cross-cultural validity, criterion validity and responsiveness were considered (**Appendix S2**). Additionally, we included any study that evaluated any of the specified additional outcomes of the tools, including feasibility, interpretability, and desire for analgesia.

Exclusion Criteria

We excluded abstracts, editorials, reviews and studies that included paediatric or adolescent populations, or sedated, mechanically ventilated and critically ill patients.

Selection of Articles

Following our database search, we collated and uploaded all identified citations to EndNote X9 (Clarivate Analytics, Philadelphia, PA, USA) and removed duplicates. The identified studies were uploaded to Rayyan QCRI online software.[25] Two reviewers (RMB and AI) independently applied the inclusion criteria to the titles, then to relevant abstracts. Afterwards, we thoroughly examined potentially eligible full texts for inclusion. We documented the full search results in the PRISMA flow diagram (**Figure 1**). Excluded studies and the reasons for their exclusion are provided in **Appendix S3**.

Data Extraction

One reviewer (RMB) extracted data from the included full-text articles, with the extraction verified by a second reviewer (AI). The two reviewers resolved any disagreements through discussion, or consultation with other reviewers (RDK, LST or DNL) when necessary. The data extracted included specific details about the assessment tool used, country, language of scale administration, study design, patient characteristics, surgical procedure, the specific measurement properties assessed, outcomes related to the review question and objectives, and the main statistical analysis.

Assessment of Methodology

Two independent reviewers (RMB and AI) critically appraised the methodological quality of studies looking at feasibility and interpretability using a modified version of the Newcastle

Ottawa Scale[26] (**Appendix S4**). For validation studies, we assessed the quality using the COSMIN criteria for methodological quality.[27-29] We included three phases in the assessment of each measurement property. First, we assessed the risk of bias which pertains to methodological quality in each study: very good, adequate, doubtful, or inadequate quality was assigned to each study. Second, we related the results to a measurement property rated against criteria for "sufficient measurement properties" and the results were classified as sufficient, insufficient, or indeterminate (**Appendix S5**). Third, we combined the results from each study and graded the quality of evidence for each pain assessment tool. A summary of the scoring criteria and appraisals is provided in (**Appendices S6 and S7**).

Protocol Registration

The protocol was registered (No. CRD42020213495) with the PROSPERO database and can be accessed at https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=213495.

RESULTS

The search identified 14,216 potential studies following removal of duplicates. After reviewing the titles, we excluded 13,798 for irrelevance and another 380 after abstract screening. Of the 38 remaining studies, we excluded 19 after examination of the full texts against the inclusion criteria (**Appendix S2**). An additional 12 studies were identified through searching the bibliography of eligible studies, so a total of 31 studies[2, 3, 6, 13, 30-56] (**Figure 1**) with 12498 participants were included. The number of participants in individual studies ranged from 35[30] to 3045.[31]

The distribution of male and female participants in the studies varied, with some studies including only female participants[30] or only male participants[40] and others not reporting sex distribution.[38 , 50, 52, 53] The studies matching our inclusion criteria were published between 1982[52] and 2018,[37] and assessed postoperative pain following different types of surgical procedures (**Table 1**). Nine studies included only cognitively intact[6, 32, 35, 38 , 47, 49, 51, 54, 55] while two studies included mild cognitively impaired participants.[46, 56] The remaining 20 studies did not report on cognitive function.[2, 3, 13, 30, 32-36, 39-45, 48, 50, 52, 53]

Seven studies were performed in the USA,[3, 36-38, 44, 45, 52] three in China,[46 , 47, 56] three in Australia,[48-50] and two each in the UK,[35, 43] Netherlands,[13, 54] Ghana,[33, 42] France[32] and Canada.[6, 40] One study each was performed in Finland,[51] Spain,[34] Nigeria,[30] Iran,[39] India,[53] Vietnam,[55] Israel,[2] and Germany.[41] Although all the included studies were reported in English, some of the tools were administered in other languages: Chinese,[46, 47, 56] Twi,[33, 42] Vietnamese,[55] Finnish,[51] and both English and Yoruba.[30]

Using the modified Newcastle Ottawa Score, the majority of studies looking at feasibility were of medium[2, 30, 32, 33, 37, 39, 49, 54] or high quality.[3, 6, 13, 35 , 36, 41, 46-48, 50, 51] The methodological quality of three secondary analysis studies that looked at VAS interpretability could not be assessed.[44, 45, 52] The methodological quality for other measurement properties is described under each measurement property section.

The following measurement properties were assessed: measurement error (n=1),[37] cross-cultural validity (n=1),[42] reliability (n=5),[33, 46-48, 56] responsiveness (n=7)[33, 40, 43, 45-47, 55] and hypothesis testing for construct validity (namely convergent validity; n=13)[6, 30, 33-35, 38-40, 46, 47, 54-56] and criterion validity (n=2).[6, 56] No studies assessed structural validity, internal consistency, or content validity of any pain assessment tool. Interpretability was measured in eleven studies.[2, 3, 31, 36, 41, 44, 48-50, 52, 54] Two studies included the desire for analgesics as an outcome.[3 , 13] The feasibility of pain assessment tools as an outcome measure was examined in eight studies.[6, 32, 33, 35, 46, 47, 51, 56]

Outcomes for measurement properties

1. Unidimensional pain assessment tools

Convergent validity

Eight studies[6, 30, 34, 35, 38-40, 47] reported the convergent validity of the VAS with moderate-to-high correlations between several self-report scales that also measured pain intensity. Similarly, seven studies reported good convergent validity results for VRS,[6, 34, 35, 45, 47, 54, 56] and six studies each reported good convergent validity results for NRS[6, 33, 46, 47, 54, 56] and

FPS[33, 39, 46, 47, 55, 56] scores (**Table 2**). The correlations between scores obtained from several unidimensional tools were moderate to high, ranging from 0.5 to 0.9.

Cross-cultural validity

One study[42] established the validity of a Twi (Ghanaian) version of the VAS. The pain scores reported by patients using the new instrument correlated significantly with those reported by patients using the original (English) version of the VAS, with the highest correlation on the fifth postoperative day. Because of inadequate quality due to an extremely serious risk of bias and imprecision, very low-quality evidence was reported for cross-cultural validity of the VAS.

Reliability

The VAS showed high scale,[46, 47] and test-retest reliability[48] with an intraclass correlation coefficient of 0.79 (95% CI: 0.49 to 0.91).[48] The NRS demonstrated high test-retest,[56] inter-rater[44] and scale reliability.[33, 46, 47, 56] VDS demonstrated high scale[47] and test-retest reliability.[56] Similarly, FPS demonstrated high inter-rater[33] and test-retest reliability[56] (**Table 3**). All four scales showed low-quality evidence due to very serious risk of bias.

Responsiveness

Seven studies[33, 40, 43, 45-47, 55] reported responsiveness results for the four unidimensional pain assessment tools and provided low-quality evidence due to a very serious risk of bias (**Table 4**). The identified risk of bias was mainly related to the use of inappropriate measures of responsiveness like effect size and statistical tests used.

Measurement error

Only one study assessed measurement error of VAS by determining the minimal detectable change (MDC),[37] which describes the smallest change outside of inherent measurement error that the VAS can detect. The study showed that the MDC on a 100 mm VAS was 15 mm for total hip arthroplasty and 16 mm for total knee arthroplasty.[37] We evaluated the evidence regarding VAS measurement error as moderate-quality because we could not determine the minimal important change for VAS in acute pain to compare with MDC and the risk of bias.

2. Functional pain assessment tool

Only one study examined the 'Objective Pain Score' which assesses the interference of pain with respiratory function.[53] The study evaluated the correlation between scores obtained from Objective Pain Score and NRS. While patients rated their pain using a printed NRS, the clinician rated pain using the Objective Pain Score. A linear regression model determined the relationship between NRS and Objective Pain Score and showed that for every unit increase in the NRS, the Objective Pain Score decreased by 0.334. The study reported sufficient convergent validity with the NRS, although with low-quality evidence due to risk of bias and imprecision. A summary of finding on all assessed measurement properties is provided in (**Table 2**).

Other outcomes

Interpretability and desire for analgesics

Visual analogue scale (VAS)

Seven stuidies[31, 37, 44, 48-50, 52] looked at the interpretability of VAS, and one study[3] included the desire for analgesics as an outcome. Several studies[31, 44, 52] reported nearly similar cut-off points for VAS, indicating that VAS ratings of 0-5 mm were very likely to be rated as no pain by patients, 6-44 mm were considered mild pain, 45-69 mm were considered moderate pain, and VAS ratings ≥70 mm were suggestive of severe pain.

Two studies[37, 48] determined the interpretability of VAS by identifying the minimal clinically-important difference (MCID) defined as the minimal change in score indicating a meaningful change in pain status.[57] The use of a combination of distribution- and anchor-based methods resulted in an MCID of 9.9 mm for VAS in assessing several types of surgical procedures.[48] In contrast, Danoff *et al*.[37] reported higher MCID values for pain improvement in patients undergoing total hip or knee arthroplasty. Pain was improving clinically when the VAS decreased by 19 and 23 mm, respectively.

Bodian *et al*.[3] found that the proportion of patients requesting additional analgesia following abdominal surgery increased as VAS increased (4%, 43%, and 80% with VAS scores of 30 mm or less, 31-70 mm, and greater than 70 mm, respectively).

Numerical rating scale (NRS)

Four studies[2, 36, 41, 54] looked at interpretability of the NRS, one study include desire for analgesics as an outcome.[13] Sloman *et al*.[2] determined the meaning of changes in NRS in relation to perceived pain relief before and after treatment. Patients who rated their pain relief as ‘minimal’ had, on average, a 35% reduction in NRS. NRS was less sensitive to detect changes from ‘moderate’ to ‘much’ as there was a 67% reduction for those who rated their reduction as ‘moderate’, a 70% decrease for those who rated it is as ‘much’, and a 94% reduction for those assessed their pain reduction as ‘complete’.[2]

Inconsistent cut-off points between moderate to severe pain were identified for NRS. For example, Gerbershagen *et al*.[41] determined NRS ≥4 as a cut-point for moderate pain, while ‘pain interfering with function’ resulted in a lower cut-off point of NRS ≥3. While using receiver operating characteristic analysis in another study, Van Dijk *et al*.[54] found that the sensitivity of NRS to differentiate bearable pain (VRS £2) from unbearable pain (VRS >2) reached higher values (94%) for high cut-off point of NRS >5 compared with lower cut-off points of 3 and 4 (sensitivity 72%, 83%) respectively.

In another study, Van Dijk *et al*.[13] showed that 19% of patients with NRS scores ranging from 5-10 had no desire for additional opioids; 62% reported that they did not want additional opioids because their pain was tolerable. When patients were asked at which score, they would request opioids, both the median and the modal pain scores were an NRS of 8.

Feasibility

Eight studies included feasibility of pain assessment tools as an outcome measure.[6, 32, 33, 35, 46, 47, 51, 56] Error rates were reported as an inability to understand the tool, responses that could not be scored reliably, and lack of responses.[6, 35, 47, 51] Some studies reported the most preferred scale or the easiest to complete ones.[6, 33, 46, 56] There was a lack of studies that assessed the time required to complete the tool or time taken to train patients or nurses.

For multiple types of surgical procedures and in different populations VDS or VRS were more successful when compared with other tools. Using VRS in patients aged ≥75 years after cardiac surgery showed a higher success rate (81%) compared with VAS (60%) and the FPS (44%). These rates varied significantly on all postoperative days ($P < 0.02$).[51] The reported reasons for the failure rate, which was identified as failure to understand or express level of pain using the assessment tool, were postoperative confusion, delirium, exhaustion, and an inability to differentiate between facial expressions.[51] In a similar way, VRS was more suited for compliance and ease of use following orthopaedic surgery compared with VAS in which 56% of patients included in the study did not understand how to complete VAS and one-third could not perform the assessment using VAS due to visual or hearing impairment.[35] Moreover, VAS showed the highest error rate of 12.3% when used in Chinese populations, whereas VRS reported the lowest error rate (0.8%), which was statistically significant ($P < 0.05$).[47] Interestingly, 40% of the patients rated NRS as the easiest, most preferred tool for assessment; on the contrary, VAS was reported the least preferred.[6]

From the nurses' perspectives in post-anaesthesia care units, NRS was the most preferred tool in 60% of the included sample.[32] Even though the VAS was the recommended tool to be used in the institution where the study was conducted, 50% of the nurses preferred to use either NRS or VRS due its complexities making it difficult for patients to understand VAS.[32] Three studies reported FPS as the preferred tool among a Chinese population[47], for women[46], middle-aged adults, and elderly patients without and with mild cognitive impairment, followed by VRS and NRS.[56] Likewise, FPS (55%) was preferred to NRS (33%) among a Ghanaian population.[33]

DISCUSSION

This systematic review presents a comprehensive examination of the measurement properties of unidimensional and functional assessment tools used for adult postoperative patients. The quality of evidence for the measurement properties and utility of the VAS, VDS, NRS, and FPS was suboptimal. Overall, construct validity (convergent validity) was most commonly assessed across measures. Content validity, internal consistency and structural validity were not assessed as these measures are not designed for single-item scales. The VAS had the greatest number of studies assessing its measurement properties in the postoperative setting, followed by the NRS. Studies on functional pain assessment tools were scarce. Most of the reviewed studies failed to meet the COSMIN methodological standards required. Good-quality studies were found for interpretability and feasibility as assessed by the Newcastle Ottawa Scale.[26]

Most of the studies reported sufficient convergent validity of several unidimensional pain assessment tools, indicating that the scales tended to measure score variations in the same direction.[58] Similar positive findings of good convergent validity results were reported when these tools were used to assess pain associated with rheumatoid arthritis[59] and osteoarthritis,[60] and low back pain.[61] However, the methodology used to measure convergent validity was limited. Because no gold standard tool exists for assessing pain, most studies assessed the correlation of scores obtained from one unidimensional tool with another, measuring only pain intensity. However, when a multidimensional tool such as the McGill Pain Questionnaire was used as a comparator, studies reported lower correlation scores.[6, 40, 62] This variation may be related to assessor and patient fatigue during the detailed pain assessment.

There was good reliability of pain assessment for all the unidimensional tools. However, the quality of evidence was low for all four scales because of serious risk of bias due to unreported intervals for repeated measures or the use of inappropriate reliability measures by treating ranked NRS, VDS or FPS scores as a continuous value. Measurement error was only available for VAS; however, the study outcome was indeterminate because we could not determine for VAS in acute pain to compare it with the MDC. When the MDC is smaller than the minimal important change, significant change can be distinguished from measurement error.[63]

Small, albeit statistically significant changes in VAS do not necessarily indicate clinically important changes to guide the interpretation of studies evaluating analgesic therapies.[37] Therefore, obtaining an accurate MCID is crucial.[64] Previous studies have shown that the MCID differs by patient population and diagnosis. We identified two studies reporting inconsistent MCID values for the postoperative population.[37, 48] The MCID tended to be higher in patients who underwent joint arthroplasties than other procedures.[48] One explanation might be that patients reporting severe, acute pain need a larger reduction in pain to be clinically meaningful.[65]

Measures of responsiveness are an important psychometric property to assess the sensitivity of change in pain over time.[66] Measures of responsiveness used included effect size, standardized response mean and scores pre- and post-intervention.[33, 40, 43, 44, 46, 47, 55] According to COSMIN methodology, effect size and standardized response mean are inappropriate to assess responsiveness because they measure the size of the change scores rather than their validity. Moreover, the *P* value of statistical tests only measures the statistical significance of the change in scores rather than their validity.[63]

Pain assessment tools help diagnose surgical catastrophes, allow communication between health care providers, and are used to assess efficacy of analgesic treatments and allow comparison between therapies. As no agreement exists on how to identify the optimal cut-off point of a unidimensional pain assessment tool, various arbitrarily chosen values are used.[41] Generally, VAS cut-off points of 30, 70, 100 mm indicate the upper boundaries of mild, moderate and severe pain. However, a recent study conducted found a higher cut- off point between mild and moderate pain of around 55 mm on the VAS, which is greater than the values reported by most earlier studies and physicians' consensus.[44, 67-69]

NRS cut-off points used by healthcare professionals do not necessarily reflect patients' desire for additional analgesics.[13] Previous studies have also found that a high proportion of patients with pain scores >4 did not demand analgesics (28% of patients visiting an emergency department[70] and 42% of children after surgery[71]). Cho *et al*.[62] showed that postoperative patients requested an analgesic when their pain was VAS ≥5.5, NRS ≥6, FPS-R ≥6 or VRS ≥2 (moderate or severe pain). This might be influenced by a general refusal for analgesic medicines, or fear of side effects or addiction, especially with opioids.[13, 72, 73] Cut-off points, although important are not validated to guide analgesic interventions.

Previously, postoperative pain assessment and management was focused on providing humanitarian pain relief, which constitutes only one objective to tackle a complex experience, and that was achieved by using unidimensional scores. However, health care providers should address pain by several approaches to determine if the pain is tolerable, is hindering recovery or requires intervention.[62]

Efforts have been made to encourage use of multidimensional tools to assess postoperative pain. A recent systematic review indicated that the Brief Pain Inventory and the American

Pain Society Pain Outcomes Questionnaire – Revised were the two commonly used and studied multidimensional pain assessment tools for patients after surgery, followed by the McGill Pain Questionnaire. These multidimensional tools showed good ratings for some psychometric properties like internal consistency. However, this recommendation was based on low- to moderate-quality evidence.[66] Moreover, these tools involve a detailed assessment that can range from 5 to 30 minutes,[74]hindering routine use for frequent assessment in a busy surgical ward.[20] Alternatively, functional pain assessment has been recommended.[14, 75]

However, since no gold standard objective measures exist for pain-related functional capacity in postoperative patients[76], we included objective tools assessing the impact of pain on function. Only one study reported sufficient convergent validity of functional assessment based on pain interference with normal breathing and NRS score.[53] The low methodological quality of the study limits the generalizability of the result. Other researchers have tried to incorporate a non- formally validated three-level 'Functional Activity Score'[20] into clinical practice. One study of a Chinese population combining the Functional Activity Score and dynamic NRS found that this allowed nurses to guide and educate patients to better use patient-controlled analgesia to facilitate functional recovery.[77] Additionally, a pilot study with hospitalized patients validated a four-level scale (no interference, interference with some or most activities, or inability to do any activity).[78] It established the convergent validity of this tool compared with NRS and VAS in cognitively intact patients. Patients aged ≥40 years also preferred a functional assessment scale,[78] possibly because functional assessment considered the impact of pain on activity.

The heterogeneity of study designs, including the assessment scales used, surgical procedures, sample sizes, countries in which the studies were conducted, and the languages used, make determining the most feasible assessment tool difficult. However, the VAS showed the highest error rate and was the least preferred in several studies, whereas the VRS showed the lowest error rate. Difficulties comprehending the VAS and linearly quantifying pain resulted in a higher frequency of incomplete responses, especially for older patients.[12, 13] Therefore, older adults and children who have less abstract thinking ability might prefer a categorical scale like the VRS for easier use.[14] Interestingly, although the FPS is commonly used in paediatric populations, it was also the most preferred tool in the Ghanaian and Chinese adult populations. This might be because of the simplicity of facial expressions, which can quickly reflect pain. Alternatively, cultural aspects may explain why the FPS was preferred.[79]

Strengths and Limitations

The main strength of this review is that it includes the most frequently used unidimensional and functional pain assessment tools. In addition, we put no limits on publication date, enabling us to obtain information on early studies of these tools. To our knowledge, this is the first review to evaluate the validity of these tools focusing solely on postsurgical populations and applying COSMIN methodology.

Potential limitations include the fact that the search strategy may have excluded grey literature and studies published in languages other than English. However, we tried to limit the effect of language and publication biases by searching the references of included studies.

In addition, the clinical diversity and limitations in the methodologies and quality of the included studies, may have reduced the strength of the conclusions.

Conclusion

This systematic review challenges the validity and reliability of unidimensional tools to quantify pain in adult patients after surgery. Despite their extensive use, no evidence clearly suggests that one tool has superior measurement properties in assessing postoperative pain. Therefore, future studies should be prioritized to assess their validity, reliability, measurement error, and responsiveness using COSMIN methodology. Moreover, adequate quality head-to-head comparison studies are required to assess several unidimensional pain assessment tools alongside other tools covering multiple dimensions of the pain experience. In addition, because promoting function is a crucial perioperative goal, psychometric validation studies of functional pain assessment tools are needed to identify patients who need additional interventions to promote recovery and improve postoperative pain assessment and management.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contribution of Dr Douglas Grindlay, Information Specialist in Centre of Evidence Based Dermatology, the University of Nottingham for his guidance in building the search strategy.

AUTHOR CONTRIBUTIONS

Study design: RMB, AI, DNL, RDK, NAL, LST

Literature search: RMB, AI

Data extraction: RMB, AI

Data analysis: RMB, AI

Data interpretation: RMB, AI, DNL, RDK, NAL, LST

Writing of the manuscript: RMB, AI, DNL, RDK, NAL, LST

Critical review: RMB, AI, DNL, RDK, NAL, LST

Approval of submitted manuscript: RMB, AI, DNL, RDK, NAL, LST

Overall supervision: DNL, RDK

COMPETING INTERESTS

None of the authors has a conflict of interest to declare

FUNDING

This work was supported by the Medical Research Council [grant number MR/K00414X/1]; and Arthritis Research UK [grant number 19891]. The PhD studentship for RMB was funded by the Saudi Cultural Bureau and the University of Nottingham.

REFERENCES

1 Carr DB, Goudas LC. Acute pain. *Lancet* 1999; **353**: 2051-8

2 Sloman R, Wruble AW, Rosen G, Rom M. Determination of clinically meaningful levels of pain reduction in patients experiencing acute postoperative pain. *Pain Manag Nurs* 2006; **7**: 153-8

3 Bodian CA, Freedman G, Hossain S, Eisenkraft JB, Beilin Y. The visual analog scale for pain. *Anesthesiology* 2001; **95**: 1356-61

4 Breivik H, Borchgrevink PC, Allen SM, et al. Assessment of pain. *Br J Anaesth* 2008; **101**: 17-24

5 Ravaud P, Keita H, Porcher R, Durand-Stocco C, Desmonts J, Mantz J. Randomized clinical trial to assess the effect of an educational programme designed to improve nurses' assessment and recording of postoperative pain. *Br J Surg* 2004; **91**: 692-8

6 Gagliese L, Weizblit N, Ellis W, Chan VWS. The measurement of postoperative pain: a comparison of intensity scales in younger and older surgical patients. *Pain* 2005; **117**: 412-20

7 Joyce C, Zutshi D, Hrubes V, Mason R. Comparison of fixed interval and visual analogue scales for rating chronic pain. *Eur J Clin Pharmacol* 1975; **8**: 415-20

8 Jensen MP, Karoly P, Braver S. The measurement of clinical pain intensity: a comparison of six methods. *Pain* 1986; **27**: 117-26

9 Ohnhaus EE, Adler R. Methodological problems in the measurement of pain: a comparison between the verbal rating scale and the visual analogue scale. *Pain* 1975; **1**: 379-84

10 Le Resche L, Burgess J, Dworkin S. Reliability of visual analog and verbal descriptor scales for" objective" measurement of temporomandibular disorder pain. *J Dent Res* 1988; **67**: 33-6

11 Wong DL, Baker CM. Pain in children: comparison of assessment scales. *Pediatr Nurs* 1988; **14**: 9-17

12 Coll AM, Ameen JR, Mead D. Postoperative pain assessment tools in day surgery: literature review. *J Adv Nurs* 2004; **46**: 124-33

13 Van Dijk JFM, Kappen TH, Schuurmans MJ, van Wijck AJM. The relation between patients' NRS pain scores and their desire for additional opioids after surgery. *Pain Pract* 2015; **15**: 604-9

14 Pasero C, Quinlan-Colwell A, Rae D, Broglio K, Drew D. American Society for Pain Management Nursing position statement: prescribing and administering opioid doses based solely on pain intensity. *Pain Manag Nurs* 2016; **17**: 170-80

15 Chou R, Gordon DB, de Leon-Casasola OA, et al. Management of postoperative pain: a clinical practice guideline from the American pain society, the American Society of Regional Anesthesia and Pain Medicine, and the American Society of Anesthesiologists' committee on regional anesthesia, executive committee, and administrative council. *J Pain* 2016; **17**: 131-57

16 Mularski RA, White-Chu F, Overbay D, Miller L, Asch SM, Ganzini L. Measuring pain as the 5th vital sign does not improve quality of pain management. *J Gen Intern Med* 2006; **21**: 607-12

17 Frasco PE, Sprung J, Trentman TL. The impact of the Joint Commission for Accreditation of Healthcare Organizations pain initiative on perioperative opiate consumption and recovery room length of stay. *Anesth Analg* 2005; **100**: 162-8

18 Vila H, Smith RA, Augustyniak MJ, et al. The efficacy and safety of pain management before and after implementation of hospital-wide pain management standards: is patient safety compromised by treatment based solely on numerical pain ratings? *Anesth Analg* 2005; **101**: 474-80

19 Laycock HC, Harrop-Griffiths W. Assessing pain: how and why? *Anaesthesia* 2021; **76**: 559-62

20 Levy N, Sturgess J, Mills P. “Pain as the fifth vital sign” and dependence on the “numerical pain scale” is being abandoned in the US: why? *Br J Anaesth* 2018; **120**: 435-8

21 Levy N, Mills P, Rockett M. Post-surgical pain management: time for a paradigm shift. *Br J Anaesth* 2019; **123**: e182-6

22 Kehlet H. Postoperative pain relief—what is the issue? *Br J Anaesth* 1994; **72**: 375-8

23 Van Boekel RLM, Vissers KCP, van der Sande R, Bronkhorst E, Lerou JGC, Steegers MAH. Moving beyond pain scores: multidimensional pain assessment is essential for adequate pain management after surgery. *PLoS One* 2017; **12**: e0177345

24 Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol* 2009; **62**: e1-e34

25 Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev* 2016; **5**: 210

26 GA Wells BS, D O'Connell, J Peterson, V Welch, M Losos, P Tugwell. The Newcastle-Ottawa Scale (NOS) for assessing the quality of non randomised studies in meta-analyses.

27 Prinsen CA, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018; **27**: 1147-57

28 Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2012; **21**: 651-7

29 Mokkink LB, De Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res* 2018; **27**: 1171-9

30 Akinpelu AO, Olowe OO. Correlative study of 3 pain rating scales among obstetric patients. *Afr J Med Med Sci* 2002; **31**: 123-6

31 Aubrun F, Langeron O, Quesnel C, Coriat P, Riou B. Relationships between measurement of pain using visual analog score and morphine requirements during postoperative intravenous morphine titration. *Anesthesiology* 2003; **98**: 1415-21

32 Aubrun F, Paqueron X, Langeron O, Coriat P, Riou B. What pain scales do nurses use in the postanaesthesia care unit? *Eur J Anaesthesiol* 2003; **20**: 745-9

33 Aziato L, Dedey F, Marfo K, Avoka Asamani J, Clegg-Lamptey JNA. Validation of three pain scales among adult postoperative patients in Ghana. *BMC Nurs* 2015; **14**: 42

34 Banos JE, Bosch F, Canellas M, Bassols A, Ortega F, Bigorra J. Acceptability of visual analogue scales in the clinical setting: a comparison with verbal rating scales in postoperative pain. *Methods Find Exp Clin Pharmacol* 1989; **11**: 123-7

35 Briggs M, Closs JS. A descriptive study of the use of visual analogue scales and verbal rating scales for the assessment of postoperative pain in orthopedic patients. *J Pain Symptom Manage* 1999; **18**: 438-46

36 Cepeda MS, Africano JM, Polo R, Alcala R, Carr DB. What decline in pain intensity is meaningful to patients with acute pain? *Pain* 2003; **105**: 151-7

37 Danoff JR, Goel R, Sutton R, Maltenfort MG, Austin MS. How much pain is significant? Defining the minimal clinically important difference for the visual analog scale for pain after total joint arthroplasty. *J Arthroplasty* 2018; **33**: S71-5. e2

38 Deloach LJ, Higgins MS, Caplan AB, Stiff JL. The visual analog scale in the immediate postoperative period. *Anesth Analg* 1998; **86**: 102-6

39 Fadaizadeh L, Emami H, Samii K. Comparison of visual analogue scale and faces rating scale in measuring acute postoperative pain. *Arch Iran Med* 2009; **12**: 73-5

40 Gagliese L, Katz J. Age differences in postoperative pain are scale dependent: a comparison of measures of pain intensity and quality in younger and older surgical patients. *Pain* 2003; **103**: 11-20

41 Gerbershagen HJ, Rothaug J, Kalkman CJ, Meissner W. Determination of moderate-to-severe postoperative pain on the numeric rating scale: a cut-off point analysis applying four different methods. *Br J Anaesth* 2011; **107**: 619-26

42 Hamzat T, Samir M, Peters G. Development and some psychometric properties of Twi (Ghanaian) version of the visual analogue scale. *Afr J Biomed Res* 2009; **12**: 145-8

43 Jenkinson C, Carroll D, Egerton M, et al. Comparison of the sensitivity to change of long and short form pain measures. *Qual Life Res* 1995; **4**: 353-7

44 Jensen MP, Chen C, Brugger AM. Interpretation of visual analog scale ratings and change scores: a reanalysis of two clinical trials of postoperative pain. *J Pain* 2003; **4**: 407-14

45 Jensen MP, Chen C, Brugger AM. Postsurgical pain outcome assessment. *Pain* 2002; **99**: 101-9

46 Li L, Herr K, Chen P. Postoperative pain assessment with three intensity scales in Chinese elders. *J Nurs Scholarsh* 2009; **41**: 241-9

47 Li L, Liu X, Herr K. Postoperative pain intensity assessment: a comparison of four scales in Chinese adults. *Pain Med* 2007; **8**: 223-34

48 Myles P, Myles D, Galagher W, et al. Measuring acute postoperative pain using the visual analog scale: the minimal clinically important difference and patient acceptable symptom state. *Br J Anaesth* 2017; **118**: 424-9

49 Myles PS, Troedel S, Boquest M, Reeves M. The pain visual analog scale: is it linear or nonlinear? *Anesth Analg* 1999; **89**: 1517-20

50 Myles PS, Urquhart N. The linearity of the visual analogue scale in patients with severe acute pain. *Anaesth Intensive Care* 2005; **33**: 54-8

51 Pesonen A, Suojaranta-Ylinen R, Tarkkila P, Rosenberg PH. Applicability of tools to assess pain in elderly patients after cardiac surgery. *Acta Anaesthesiol Scand* 2008; **52**: 267-73

52 Sriwatanakul K, Kelvie W, Lasagna L, Calimlim JF, Weis OF, Mehta G. Studies with different types of visual analog scales for measurement of pain. *Clin Pharmacol Ther* 1983; **34**: 234-9

53 Tandon M, Singh A, Saluja V, Dhankhar M, Pandey CK, Jain P. Validation of a new “Objective Pain Score” Vs. “Numeric Rating Scale” for the evaluation of acute pain: a comparative study. *Anesth Pain Med* 2016; **6**: e32101

54 Van Dijk JF, Kappen TH, van Wijck AJ, Kalkman CJ, Schuurmans MJ. The diagnostic value of the numeric pain rating scale in older postoperative patients. *J Clin Nurs* 2012; **21**: 3018-24

55 Van Giang N, Chiu H-Y, Thai DH, Kuo S-Y, Tsai P-S. Validity, sensitivity, and responsiveness of the 11-face faces pain scale to postoperative pain in adult orthopedic surgery patients. *Pain Manag Nurs* 2015; **16**: 678-84

56 Zhou Y, Petpichetchian W, Kitrungrote L. Psychometric properties of pain intensity scales comparing among postoperative adult patients, elderly patients without and with mild cognitive impairment in China. *Int J Nurs Stud* 2011; **48**: 449-57

57 Farrar JT, Young Jr JP, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* 2001; **94**: 149-58

58 Hjermstad MJ, Fayers PM, Haugen DF, et al. Studies comparing 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

59 Sendlbeck M, Araujo EG, Schett G, Englbrecht M. Psychometric properties of three single-item pain scales in patients with rheumatoid arthritis seen during routine clinical care: a comparative perspective on construct validity, reproducibility and internal responsiveness. *RMD Open* 2015; **1**: e000140

60 Alghadir AH, Anwer S, Iqbal A, Iqbal ZA. Test-retest reliability, validity, and minimum detectable change of visual analog, numerical rating, and verbal rating scales for measurement of osteoarthritic knee pain. *J Pain Res* 2018; **11**: 851-6

61 Chiarotto A, Maxwell LJ, Ostelo RW, Boers M, Tugwell P, Terwee CB. Measurement properties of visual analogue scale, numeric rating scale, and pain severity subscale of the brief pain inventory in patients with low back pain: a systematic review. *J Pain* 2019; **20**: 245-63

62 Cho S, Kim YJ, Lee M, Woo JH, Lee HJ. Cut-off points between pain intensities of the postoperative pain using receiver operating characteristic (ROC) curves. *BMC Anesthesiol* 2021; **21**: 29

63 De Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine: A Practical Guide. Cambridge: Cambridge University Press, 2011

64 Wells G, Beaton D, Shea B, et al. Minimal clinically important differences: review of methods. *J Rheumatol* 2001; **28**: 406-12

65 Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. *Ann Rheum Dis* 2005; **64**: 34-7

66 Lapkin S, Ellwood L, Diwan A, Fernandez R. Reliability, validity, and responsiveness of multidimensional pain assessment tools used in postoperative adult patients: a systematic review of measurement properties. *JBI Evid Synth* 2021; **19**: 284-307

67 Boonstra AM, Preuper HRS, Balk GA, Stewart RE. Cut-off points for mild, moderate, and severe pain on the visual analogue scale for pain in patients with chronic musculoskeletal pain. *Pain* 2014; **155**: 2545-50

68 Serlin RC, Mendoza TR, Nakamura Y, Edwards KR, Cleeland CS. When is cancer pain mild, moderate or severe? Grading pain severity by its interference with function. *Pain* 1995; **61**: 277-84

69 Zelman DC, Hoffman DL, Seifeldin R, Dukes EM. Development of a metric for a day of manageable pain control: derivation of pain severity cut-points for low back pain and osteoarthritis. *Pain* 2003; **106**: 35-42

70 Blumstein HA, Moore D. Visual analog pain scores do not define desire for analgesia in patients with acute pain. *Acad Emerg Med* 2003; **10**: 211-4

71 Voepel-Lewis T, Burke CN, Jeffreys N, Malviya S, Tait AR. Do 0–10 numeric rating scores translate into clinically meaningful pain measures for children? *Anesth Analg* 2011; **112**: 415-21

72 Gan T, Lubarsky D, Flood E, et al. Patient preferences for acute pain treatment. *Br J Anaesth* 2004; **92**: 681-8

73 Miaskowski C. The impact of age on a patient's perception of pain and ways it can be managed. *Pain Manag Nurs* 2000; **1 (3 Suppl 1)**: 2-7

74 Wilkie DJ, Savedra MC, Holzemer WL, Tesler MD, Paul SM. Use of the McGill Pain Questionnaire to measure pain: a meta-analysis. *Nurs Res* 1990; **39**: 36-41

75 Levy N, Quinlan J, El-Boghdadly K, et al. An international multidisciplinary consensus statement on the prevention of opioid-related harm in adult surgical patients. *Anaesthesia* 2021; **76**: 520-36

76 White PF, Kehlet H. Improving postoperative pain management: what are the unresolved issues? *Anesthesiology* 2010; **112**: 220-5

77 Tong YG, Konstantatos AH, Yan C, Ling C. Improving pain management through addition of the functional activity score. *Aust J Adv Nurs* 2018; **35**: 52-60

78 Halm M, Bailey C, St Pierre J, et al. Pilot evaluation of a functional pain assessment scale. *Clin Nurse Spec* 2019; **33**: 12-21

79 Pasero C, McCaffery M. Pain Assessment and Pharmacologic Management. St. Louis: Mosby, 2010

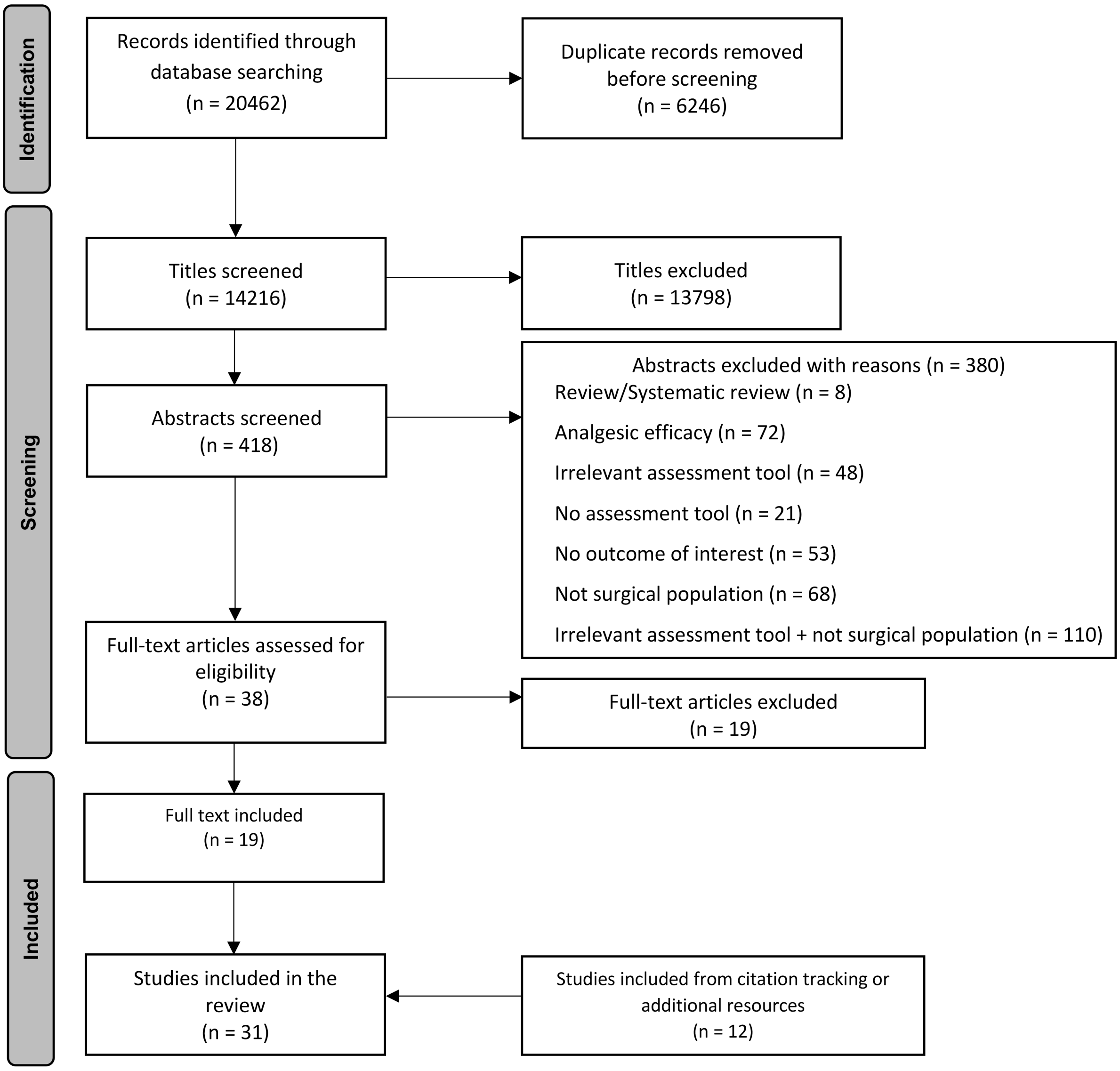
Identification
Records identified through database searching (n = 20462)
Duplicate records removed before screening (n = 6246)
Screening
Titles screened (n = 14216)
Titles excluded (n = 13798)
Abstracts screened (n = 418)
Abstracts excluded with reasons (n = 380)
Review/Systematic review (n = 8)
Analgesic efficacy (n = 72)
Irrelevant assessment tool (n = 48)
No assessment tool (n = 21)
No outcome of interest (n = 53)
Not surgical population (n = 68)
Irrelevant assessment tool + not surgical population (n = 110)
Full-text articles assessed for eligibility (n = 38)
Full-text articles excluded (n = 19)
Included
Full text included (n = 19)
Studies included in the review (n = 31)
Studies included from citation tracking or additional resources (n = 12)

Figure 1: PRISMA Diagram

Table 1: Characteristics of included studies

First Author Year Country	PROM/s	Study Design	Surgical Procedure	Outcome/s	High Anchor*	Main Exclusion Criteria	Patient Characteristics	
							n (Female%)	Age Years, Mean ± SD (range)
Van Dijk 2015[13] Netherlands	NRS	Cross-sectional design	Orthopaedic, ENT, gynaecological, cardiothoracic, Others	Ability to detect desire for analgesics	Worst pain imaginable	ICU patients, not proficient in Dutch or English, ambulatory surgery	1,084 (48)	53 (18–90)
Banos 1989[34] Spain	VAS VRS-5	Descriptive correlational design	Abdominal, orthopaedic, gynaecological	Convergent validity	Number 10 Unbearable pain	NR	212 (50)	<30 = 43 31-50 = 69 >50 = 107
Akinpelu 2002[30] Nigeria	VAS M-VRS BNS	Cross-sectional design	Caesarean section	Convergent validity	Worst pain Worst Imaginable Worst pain	Complications, Illness Unconscious	35 (100)	31 ± 5
Briggs 1999[35] UK	VAS VRS**	Secondary analysis of RCT	Orthopaedic	Convergent validity Feasibility	Number 100 Severe pain at rest and movement	NR	417 (45)	47 ± 20* 64 ± 17
Fadaizadeh 2009[39] Iran	VAS FPS	Cross-sectional design	General, gynaecological	Convergent validity	Number 10 Agonized 6	History of substance abuse, Unconscious	82 (72) 34 GS 48 GYN	32 ± 14 GYN 27 ± 7 GS 38 ± 18
DeLoach 1998[38] USA	VAS VPS	Descriptive correlational design	Various type of surgeries	Convergent validity	Worst imaginable Horrible pain	NR	NR	NR
Pesonen 2008[51] Finland	VAS VRS-5 RWS FPS-7	Descriptive correlational design	Cardiac surgery: elective CABG, valvular repair	Feasibility	Number 10 Unbearable pain 50 cm Number 6	Dementia, Cognitive impairment	160 FPS 80 (36) RWS 80(44)	73 ± 5
Aubrun 2003[32] France	VAS NRS VRS Behavioural scale	Prospective observational design	Orthopaedic, abdominal, gynaecological, others	Feasibility	10 worst imaginable pain VRS severe NR	NR	600 (47)	51 ± 17

First Author Year Country	PROM/s	Study Design	Surgical Procedure	Outcome/s	High Anchor*	Main Exclusion Criteria	Patient Characteristics	
							n (Female%)	Age Years, Mean ± SD (range)
Myles 1999[49] Australia	VAS	Clinical study	General, orthopaedic, ENT, faciomaxillary, cardiothoracic	Interpretability	100 worst pain ever	Severe pain, inability to complete the VAS	52 (40)	42 ± 15
Myles 2005[50] Australia	VAS	Clinical study	General, orthopaedic, ENT, faciomaxillary, cardiothoracic	Interpretability	100 worst pain ever	Postoperative delirium Frailty, visual impairment	22 (NR)	33 ± 17
Jensen 2003[44] USA	VAS VRS-4 VRS-P	Secondary analysis of RCT	Total knee replacement, hysterectomy, laparotomy	Interpretability	Worst pain Severe pain Complete relief	NR	123 (66)	65 ±10
Gerbershage 2011[41] Germany	NRS	Comparative study design	Cholecystectomy, thyroidectomy, gastrointestinal, inguinal hernia repair, others	Interpretability	Worst imaginable pain	Repeated surgical, procedures, mechanical ventilation	444 (44)	18–20 = 38 21–30 = 75 31–40 = 88 41–50 = 96 51–60 = 87 61–70 = 49 71–80 = 2
Cepeda 2003[36] USA	NRS VRS	Clinical study	Head and neck, thoracic, spinal abdominal, orthopaedic	Interpretability	Worst imaginable Severe pain	NR	700 (62)	50 ± 15
Jensen 2002[45] USA	VAS VRS Pain relief	Secondary analysis of RCT	Total knee replacement, abdominal hysterectomy, laparotomy	Responsiveness	Worst pain Severe pain Complete relief	NR	246 (66)	Knee 65 ± 10 Laparotomy 41 ± 7.5
Jenkinson 1995[43] UK	VAS CPI McGill	RCT	Orthopaedic	Responsiveness	Severe pain	NR	75 (64)	Male: 41 ± 13 Female: 43 ± 12

First Author Year Country	PROM/s	Study Design	Surgical Procedure	Outcome/s	High Anchor*	Main Exclusion Criteria	Patient Characteristics	
							n (Female%)	Age Years, Mean ± SD (range)
Aubrun 2003[31] France	VAS	Clinical study	Orthopaedic, urological, abdominal gynaecological, vascular, thoracic	Interpretability	Number 10	Minor pain, delirium, dementia, non-French speaking	3045 (54)	50 ± 18
Sriwatanakul 1982[52] USA	VAS	Secondary analysis of RCT	NR	Interpretability	Pain as bad as it could be	NR	NR	NR
Van Giang 2015[55] Vietnam	FPS NRS	Validation study	Orthopaedic	Concurrent validity Responsiveness	The worst possible pain	Hearing impairment Altered mental status	144 (45)	37 ± 13
Van Dijk 2012[54] Netherlands	NRS VRS	Cross-sectional design	General, ENT, orthopaedic, neurosurgical, urological, gynaecological, plastic, vascular, cardiothoracic	Interpretability	10 Terrible pain	ICU patients Non-Dutch speaking Cognitive or hearing impairment, inability to use self-report	2674 (51)	73 ± 6
Li 2007[47] China	VAS NRS-11 VDS FPS	Prospective clinical study	NR	Convergent validity Scale reliability Responsiveness Feasibility	10 Worst pain 10 worst pain 10 worst pain Worst pain	NR	173 (45)	45.3 ± 15
Li 2009[46] China	FPS NRS IPT	Descriptive correlational design	Gastrointestinal, orthopaedic, abdominal	Convergent validity Scale reliability Responsiveness Feasibility	10 10 The most intense imaginable pain	Did not speak Chinese More than one surgery ASA score of 4 Chronic pain	180 (68)	72 ± 6
Zhou 2011[56] China	VDS NRS FPS CAS	Descriptive comparative design	NR	Criterion validity Convergent validity Test–retest reliability Feasibility	Worst pain	Severe cognitive impairment	200 (46)	56 ± 16

First Author Year Country	PROM/s	Study Design	Surgical Procedure	Outcome/s	High Anchor*	Main Exclusion Criteria	Patient Characteristics	
							n (Female%)	Age Years, Mean ± SD (range)
Gagliese 2005[6] Canada	VAS-H VAS-V NRS VDS MPQ	Validation study	NR	Feasibility Convergent validity Criterion validity	10 Worst possible Pain 10 worst pain imaginable Excruciating	On epidural or regional analgesia, ASA score of >3 Chronic pain, Cognitive impairment, Opioid or substance abuse	504 (58)	53 ± 15
Tandon 2016[53] India	OPS NRS	Descriptive correlational design	Abdominal surgery	Convergent validity	Worst possible pain Inadequate pain relief/pain at rest	Haemodynamic instability Unable to use a PCA pump	93	NR
Aziato 2015[33] Ghana	NRS FPS CCPS	Two phases: qualitative and psychometric testing	Caesarean section, leg amputation, laminectomy, laparotomy, others	Convergent validity Inter-rater reliability Responsiveness Feasibility	Worst possible pain Hurts worst	NR	150 (77)	<30 = 44.7 30–39 = 35 40+ = 21
Hamzat 2009[42] Ghana	VAS	Validation study	Various gynaecological procedures	Cross-cultural validity	Worst possible pain	History of psychological or psychiatric disorders	60 (100)	NR
Gagliese 2003[40] Canada	MPQ PPI VAS-R VAS-M	Descriptive correlation design	Radical prostatectomy	Convergent validity Responsiveness	Worst possible pain 5 excruciating 10 worst possible 10 worst possible pain	Non–English speaker ASA >3 Chronic pain Chronic use of opioids	200	Younger patients: 56 ± 6 Older patients: 67 ± 3
Myles 2017[48] Australia	VAS	Observational design	General, orthopaedic, gynaecological, urological, major vascular, cardiac faciomaxillary, others	Test–retest reliability Interpretability	Very severe pain	Poor English comprehension Drug or alcohol dependence Psychiatric disorder Uncontrolled pain	219 (68)	53 ± 17

First Author Year Country	PROM/s	Study Design	Surgical Procedure	Outcome/s	High Anchor*	Main Exclusion Criteria	Patient Characteristics	
							n (Female%)	Age Years, Mean ± SD (range)
Danoff 2018[37] USA	VAS	Prospective observational design	THA TKA	Measurement error	Worst possible pain	Preoperative pain Catastrophising Scale score greater than 30 points	304 THA (21) TKA (30)	THA: 60 (20–81) TKA; 63 (46–88)
Sloman 2006[2] Israel	NRS	One group pretest–post-test design	Abdominal, orthopaedic, others	Interpretability	10 excruciating	NR	150 (47)	47 (14–89)
Bodian 2001[3] USA	VAS McGill	Clinical study	Intraabdominal Surgery	Interpretability Desire for analgesics	Worst pain imaginable	NR	150 (48)	49 (37–61)

PROM/s, patient-reported outcome measures; NRS, numerical rating scale; ENT, ear, nose and throat; ICU, intensive care unit; VRS-5, 5-point verbal rating scale; VAS, visual analogue scale; NR, not reported; M-VRS, modified verbal rating scale with 11 description of pain intensity; BNS, box numerical rating scale; RCT, randomized controlled trial, VRS**, four-point verbal rating scale; FPS, face pain scale; VPS, 11-point verbal scale; RWS, red wedge scale; VRS-P; verbal rating scale for pain relief; CCPS, colour circle pain scale; MPQ, McGill pain questionnaire ;VDS; verbal descriptor scale; CAS, coloured analogue scale; ASA; American Society of Anesthesiologists score; PPI, present pain intensity; OPS, objective pain score; PCA, patient controlled analgesia; VAS-R , visual analogue scale at rest, VAS-M; visual analogue scale at movement; THA, total hip arthroplasty; TKA, total knee arthroplasty.

Table 2: Summary of methodological quality of studies using COSMIN risk of bias and measurement properties

First Author	Content Validity	Structural Validity	Internal Consistency	Cross Cultural Validity	Reliability	Measurement Error	Criterion Validity	Construct Validity/ Convergent	Responsiveness
VAS			**Methodological quality assessment (COSMIN risk of bias)**						
Banos 1989[34]								Adequate	
Akinpelu 2002[30]								Doubtful	
Briggs 1999[35]								Adequate	
Fadaizadeh 2009[39]								Adequate	
DeLoach 1998[38]								Doubtful	
Li 2007[47]					Inadequate			Adequate	Inadequate
Gagliese2005[6]							Inadequate	Inadequate	
Gagliese 2003[40]								inadequate	Inadequate
Myles 2017[48]					Inadequate				
Jensen 2002[45]									Inadequate
Danoff 2018[37]						Adequate			
Hamzat 2009[42]				Inadequate					
Rating				?	+	?	?	+	?
LoE				Very low	Low	Moderate	Very low	High	Low
NRS			**Methodological quality assessment (COSMIN risk of bias)**						
Van Dijk 2012[54]								Adequate	
Li 2007[47]					Inadequate			Adequate	Inadequate
Li 2009[46]					Inadequate			Adequate	Inadequate
Zhou 2011[56]					Inadequate		Adequate	Adequate	
Gagliese 2005[6]							Inadequate	Inadequate	

Aziato 2015[33]	Inadequate		Doubtful	Inadequate
Rating	+	±	+	?
LoE	Low	low	High	Low
VDS	**Methodological quality assessment (COSMIN risk of bias)**			
Banos 1989[34]			Adequate	
Briggs 1999[35]			Adequate	
Van Dijk 2012[54]			Adequate	
Li 2007[47]	Inadequate		Adequate	
Zhou 2011[56]	Inadequate	Adequate	Adequate	
Gagliese 2005[6]		Inadequate	Inadequate	
Jensen 2002[45]				Inadequate
Rating	+	±	±	?
LoE	Low	low	High	Low
FPS	**Methodological quality assessment (COSMIN risk of bias)**			
Fadaizadeh 2009[39]			Adequate	
Van Giang 2015[55]			Adequate	Doubtful
Li 2007[47]	Inadequate		Adequate	Inadequate
Li 2009[46]	Inadequate		Adequate	Inadequate
Zhou 2011[56]	Inadequate	Adequate	Adequate	
Aziato 2015[33]	Inadequate		Doubtful	Inadequate
Rating	+	+	+	?
LoE	Low	Moderate	High	Low
OPS	**Methodological quality assessment (COSMIN risk of bias)**			
Tandon 2016[53]			Doubtful	
Rating			+	
LoE			Very low	

VAS, visual analogue scale; NRS, numerical rating scale; VDS, verbal descriptor scale; FPS, faces pain scale; OBS, objective pain score; LoE, Level of evidence using GRADE approach reported as: High, Moderate, Low, or Very low; Ratings for overall quality reported as sufficient (+), insufficient (-), inconsistent (±), indeterminate (?). Empty cells indicate no available results for measurement properties.

Table 3: Reliability of unidimensional pain assessment tools in surgical patients

First Author Year	PROM/s	Pain construct	Reliability			
			Type	n	Time interval	Interclass correlation coefficient
Li 2007[47]	VAS NRS VDS FPS	Current, worst, least, average pain on 7 postoperative days	Scale reliability	173	Every 24 hours	*0.66 *0.76 *0.72 *0.72
Li 2009[46]	FPS NRS Iowa Pain Thermometer	Current pain and daily retrospective ratings of worst and least pain	Scale reliability	180	Every 24 hours	0.95 to 0.97 ‡
Zhou 2011[56]	VDS NRS FPS Numeric Box-21 Scale Coloured Analogue Scale	Recalled pain and postoperative pain	Test–retest reliability	153	24 hours	0.96, 0.88, 0.93, 0.84¶ 0.94, 0.90, 0.91, 0.80¶ 0.93, 0.91, 0.84, 0.80¶ 0.92, 0.91, 0.78, 0.76¶ 0.93, 0.90, 0.88, 0.77¶
Aziato 2015[33]	NRS FPS Colour Circle Pain Scale	No pain – worst possible pain No pain – worst possible pain No pain – unbearable	Inter-rater reliability	150	5 to 10 minutes	0.92 0.93 0.93
Myles 2017[48]	VAS	Pain unchanged or almost the same	Test–retest reliability	22	NR	0.79 (0.49–0.91)**

PROM/s, patient-reported outcome measures; n, number of patients; VAS, visual analogue scale; NRS, numerical rating scale; VDS, verbal descriptor scale; FPS, faces pain scale; * average interclass correlation coefficient calculated for 7 days, ‡ no separate result for each scale; ¶ results categorised in 20–44 years (n = 43), 45–59 years (n = 39), 60 years without cognitive impairment (n = 40), ≥60 years with mild cognitive impairment (n = 31); ** 95% CI.

Table 4: Responsiveness results of unidimensional tools

First Author Year	PROM/s	Time Interval	n	Better, Same, Worse %	Mean Difference Pre and Post Treatment (95% CI)	Effect Size OR SRM (95% CI)	Correlation with Changes in Other Instruments
Jensen 2002[45]	VAS VDS Relief rating	Baseline then several times	123 125		10.37€, 20.71¶ 7.17€, 15.09¶ 7.59€, 26,61¶		
Jenkinson 1995[43]	VAS CPI MPQ	Baseline then 120 minutes	75	Moderate 2.23^, 1.83# Good 1.91^; 3.13# Complete 1.89^, 5#		G1;0.99^, 1.93# G2;1.23^, 1.82# G3; 2^, 3.29# G4;1.48^, 1.48#	CPI 0.67 to VAS
Van Giang 2015[55]	FPS NRS	Every 30 minutes for 2 hours	144		-1.17* -1.59+ -1.66† -1.82$	-0.70* -1.05+ -1.20† -1.31$	0.78
Li 2007[47]	VAS NRS VDS FPS	NR	28		4.3 ±2.4† 4.2 ± 2.3† 4.5 ± 2.1† 4.3 ±1.9†		
Li 2009[46]	FPS NRS IPT	NR	180		14.095 †*		
Aziato 2015[33]	NRS FPS CCPS	NR	150		2.3 (2.1–2.5) † 1.5 (1.4–1.6) † 1.4 (1.3–1.5) †		
Gagliese 2003[40]	MPQ PPI VAS-R VAS-M	NR	200			0.31¥, 0.39 0.25¥,0.26 0.23¥, 0.32 NR	

PROM/s , patient-reported outcome measures; SRM, standardized response mean; VAS, visual analogue scale; VDS, verbal descriptor scale; €, knee surgery; ¶, laparotomy surgery; ^, VAS score; #, CPI score; CPI, categorical verbal pain rating scale; MPQ, McGill pain questionnaire; G, group; FPS, face pain scale; VDS, verbal descriptor scale; FPS,

face pain scale; CCPS, colour circle pain scale; PPI, present pain intensity; VAS-R, visual analogue scale at rest; VAS-M, visual analogue scale at movement; Effect size, calculated by taking a mean change of variable and dividing it by standard deviation of that variable; *, time 2 versus time 1; +, time 3 versus time 1; †, time 4 versus time 1; $, time 5 versus time; †, p-value is statistically significant at <0.0001; ¥, results for younger patient split of the sample at the median age of 62 years.
Note: Empty cells indicate data not available or not assessed.

SUPPLEMENTARY DATA

APPENDIX S1: Search strategy

Search strategy for Ovid Medline Version 15/08/20

PICO
Population
Postoperative patients aged 18 years and over from all surgical disciplines.
Intervention
Unidimensional pain assessment tools including

◊ Verbal or printed numerical pain rating scale.

◊ Printed or verbal descriptor scale.

◊ Visual analogue scale.

◊ Faces scales: Wong-baker FACES, Faces Pain Scale – Revised.

Functional pain assessment tools
Comparison: -------
Outcomes: psychometric properties including validity and reliability
Additional outcomes
Instrument feasibility, interpretability, and ability to detect desire of analgesia.

Search concepts to be combined for Boolean AND, and used for unidimensional pain assessment tool and then repeated for functional pain assessment tools

1. Outcome terms

2. Pain assessment tool terms

3. Construct: acute postoperative pain

4. 1 AND 2 AND 3

5. 4 + Limits (english , humans, adults > 18 years)

Did not apply limits full text, abstracts this might include bias in the results
Ovid MEDLINE(R) ALL < 1946 to August 15, 2022>

1 exp PSYCHOMETRICS/ or psychometr*.mp. or measurement propert*.mp. or Validity.mp. or valid*.mp. or exp Validation Study/ or convergent validity.mp. or construct validity.mp. or content validity.mp. or criterion validity.mp. or reliab*.mp. or unreliab*.mp. or Comparative Study.mp. or Feasibility.mp. or Generalizability.mp. or generalisa*.mp. or interpretab*.mp. or Sensitiv*.mp. or Responsive*.mp. or 'Measurement Accuracy'.mp. or 'ease of use'.mp. or Analgesi* response.mp. or 'desire of analgesi*'.mp. or 'Request of

analgesic*'.mp. or 'hypotheses testing'.mp. or 'measurement error*'.mp. or Internal consistency.mp. or Data accuracy.mp. or 'standard error of measurement'.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 4890505

2 (pain scale* or pain rating scale* or (pain assessment and (instrument* or tool*)) or pain intensity scale* or pain measurement instrument* or Pain score* or pain intensity assessment).mp. or exp Pain Measurement/ [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 113996

3 Visual Analog Scale.mp. or exp Visual analog? Pain scale/ or (visual analog? and (scale or score)).mp. or vas.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 146135

4 ((numeric* and rating and (scale or score)) or numeric scale or nrs or nprs).mp. or exp numerical pain rating scale/ [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 26611

5 exp verbal descriptor scale/ or Vds.mp. or exp verbal rating scale*/ [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 1128

6 exp face* pain scale*/ or exp wong baker Face*/ or wong baker face*.mp. or exp faces pain scale revised/ or faces pain scale revised.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 594

7 (pain activity assessment or functional pain assessment scale or functional activity score*or functional pain activity scale* or functional assessment tool or objective pain score* or movement evoked pain assessment or assessment of pain at movement or objective pain assessment or clinically aligned pain assessment tool).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 252

8 exp Pain, Postoperative/ or exp acute pain/ or post surgical pain.mp. or surgical pain.mp. or pain post procedure.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 46322

9 1 and 3 and 8 5987

10 1 and 4 and 8 556

11 1 and 5 and 8 6

12	1 and 6 and 8	56
13	1 and 7 and 8	32
14	limit 9 to (Elanguage and humans and "all adult (19 plus years)")	4358
15	limit 10 to (English language and humans and "all adult (19 plus years)")	537
16	limit 11 to (English language and humans and "all adult (19 plus years)")	6
17	limit 12 to (English language and humans and "all adult (19 plus years)")	12
18	limit 13 to (English language and humans and "all adult (19 plus years)")	2

Search strategy for other databases can be provided on demand from the corresponding author

APPENDIX S2: Measurment properties included in the main domians of the COSMIN taxonomy

Domain	Psychometric property	Definition
Reliability		The extent that the measurement is free from measurement error such that scores for patients who have not changed are the same under repeated measurements
	Internal consistency	The extent that items are inter-related
	Reliability	The proportion of the total variance in the measurements that is due to ‘true’ differences between patients (as opposed to error)
	Measurement error	Error in a participant’s score that is not attributed to the construct being measured
Validity		The extent that an assessment measures what it aims to measure
	Content validity	The extent that an assessment’s content reflects the construct being measured
	Face validity	The extent that an assessment looks like it reflects the construct being measured
	Construct validity	The extent that an assessment’s scores are consistent with hypotheses based on the assumption that the tool measures what it purports to measure
	Structural validity	The extent that an assessment’s scores reflect the dimensionality of the construct being measured
	Hypothesis testing	Construct validity for the items of an assessment
	Cross-cultural validity	The extent that items on a translated or culturally modified assessment reflect the original items
	Criterion validity	The extent that an assessment’s scores represent the ‘gold standard’

Responsiveness	An assessment and/or it's items' ability to detect change over time in the construct being measured
Interpretability*	The extent that clinical or everyday understanding can be applied to an assessment's scores
Feasibility*	How easily a pain measure can be scored and interpreted

COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments. Adopted from Mokkink LB, et al.[1]

*Interpretability and *feasibility are not considered measurement properties, but important characteristics of a measurement instrument.

APPENDIX S3: Studies ineligible following full-text review

Full paper examined: 38/ Exclusion after complete paper screening 19 papers.

Excluded papers:

1. Arnstein P, Gentile D, Wilson M. validating the functional pain scale for hospitalized adults. *Pain Manag Nurs*. 2019; **20:** 418-24.

Explanation: Paper validating functional scale for hospitalized chronic pain patient but did not report separte result for surgical patients.

Reason for exclusion: No separate results for postoperative pain assessment.

2. Barber MD, Janz N, Kenton K, et al. Validation of the surgical pain scales in women undergoing pelvic reconstructive surgery. *Female Pelvic Med Reconstr Surg*. 2012; **18:** 198-204.

Explanation: Surgical pain scale looked at long term functional outcome following surgery.

Reason for exclusion: Patients not assessed as inpatients/irrelevant outcome.

3. McCarthy Jr M, Chang CH, Pickard AS, et al. Visual analog scales for assessing surgical pain. *Jl Amn Coll Surg*. 2005; **201:** 245-52.

Reason for exclusion: Patients not assessed as inpatients or irrelevant outcome.

4. Blumstein HA, Moore D. Visual analog pain scores do not define desire for analgesia in patients with acute pain. *Acad Emerg Med.* 2003; **10:** 211-4.

Explanation: VAS to detect desire of analgesia in acute emergency pain.

Reason for exclusion: Not surgical population.

5. Chiu LYL, Sun T, Ree R, et al. The evaluation of smartphone versions of the visual analogue scale and numeric rating scale as postoperative pain assessment tools: a prospective randomized trial. *Can J Anesth*. 2019; **66:** 706-15.

Reason for exclusion: Comparison between NRS smart version with paper version.

6. Neudecker J, Raue W, Schwenk W. High correlation but inadequate point-to-point agreement, between conventional mechanical and electronical visual analogue scale for assessment of acute postoperative pain after general surgery. *Acute Pain*. 2006; **8:** 175-80.

Reason for exclusion: Comparison between electronic and mechanical VAS.

7. Erden S, Karadag M, Guler Demir S, et al. Cross-cultural adaptation, validity, and reliability of the Turkish version of revised American Pain Society patient outcome questionnaire for surgical patients. *Agri*. 2018; **30:** 39-50.

Reason for exclusion: Multidimensional tool (Revised American Pain Society Patient Outcome Questionnaire).

8. Keawnantawat P, Thanasilp S, Preechawong S. Translation and validation of the Thai version of a modified brief pain inventory: a concise instrument for pain assessment in postoperative cardiac surgery. *Pain Pract*. 2017; **17:** 763-73.

Reason for exclusion: Multidimensional tool (modified brief pain inventory).

9. Mendoza TR, Chen C, Brugger A, et al. The utility and validity of the modified Brief Pain Inventory in a multiple-dose postoperative analgesic trial. *Clin J Pain*. 2004; **20:** 357-62.

Reason for exclusion: Multidimensional tool (Brief Pain Inventory).

10. Mwachiro M, Mwachiro E, Wachu M, et al. assessing post-operative pain with self-reports via the Jerrycan Pain Scale in Rural Kenya. *World J Surg*. 2020; **44:** 3636-42.

Reason for exclusion: Applicability of irrelevant tool (Jerrycan Pain Scale).

11. Jain R, Grewal A. A randomized comparative study assessing efficacy of pain versus comfort scores. *Saudi J Anaesth*. 2017; **11**: 396-401.

Reason for exclusion: Retracted paper.

12. Liu WH, Aitkenhead AR. Comparison of contemporaneous and retrospective assessment of postoperative pain using the visual analogue scale. *Br J Anaesth*. 1991; **67**: 768-71.

Reason for exclusion: Irrelevant outcome.

13. Salo D, Eget D, Lavery RF, Garner L, Bernstein S, on K. Can patients accurately read a visual analog pain scale? *Am J Emerg Med*. 2003; **21**: 515-9.

Reason of exclusion: Not surgical population.

14. Sills ES, Genton MG, Walsh APH, Wehbe SA. Who's asking? Patients may under-report postoperative pain scores to nurses (or over-report to surgeons) following surgery of the female reproductive tract. *Arch Gynecol Obstet*. 2009; **279**: 771-4.

Explanation: Looked at how patient communicate pain between nurse and physician.

Reason for exclusion: Irrelevant outcome.

15. Rothaug J, Weiss T, Meissner W. How simple can it get? Measuring pain with NRS items or binary items. *Clin J Pain*. 2013; **29**: 224-32.

Explanation: They used different answer format for (binary yes/no answers vs. NRS) in a subset of patients using Quality Improvement in Postoperative Pain Management (QUIPS).

Reason for exclusion: Multidimensional tool (QUIPS).

16. Zalon ML. Comparison of pain measures in surgical patients. *J Nurs Meas*. 1999; **7:** 135-52.

Explanation: This study aimed to establish the validity of brief pain inventory short form.

Reason for exclusion: Validation of multidimensional scale.

17. Halm M, Bailey C, St Pierre J, et al. Pilot evaluation of a functional pain assessment scale. *Clin Nurse Spec*. 2019; **33:** 12-21.

Explanation: Sample from medical/surgical, critical care, and rehabilitation units experiencing acute or chronic pain.

Reason for exclusion: No separate results for acute postoperative pain.

18. Martin WJJM, Ashton-James CE, Skorpil NE, et al. What constitutes a clinically important pain reduction in patients after third molar surgery? *Pain Res Manag*. 2013; **18:** 319-22.

Reason for exclusion: Dental surgery, not hospitalized patients.

19. Rago R, Forfori F, Materazzi G, et al. Evaluation of a preoperative pain score in response to pressure as a marker of postoperative pain and drugs consumption in surgical thyroidectomy. *Clin J Pain*. 2012; **28:** 382-6.

Reason for exclusion: Sensitivity of preoperative vas scores after tourniquet pressure inflation.

APPENDIX S4: Newcastle-Ottawa Quality Assessment Scale

(adapted for cross sectional studies)

This scale has been adapted from the Newcastle-Ottawa Quality Assessment Scale for cohort studies to perform a quality assessment of cross-sectional studies for the systematic review.

Selection: (Maximum 4 stars)

1) Representativeness of the sample:

a) Truly representative of the average in the target population. * (all subjects or random sampling)

b) Somewhat representative of the average in the target population. * (non-random sampling)

c) Selected group of users.

d) No description of the sampling strategy.

2) Sample size:

a) Justified and satisfactory. (by reporting appropriate sample size calculation) *

b) Not justified.

3) Non-respondents: (adopted to details about patient refused assessment and reasons are described)

a) Comparability between assessed and non-assessed is established *

b) The response rate is unsatisfactory, or the comparability between respondents and non-respondents is unsatisfactory. removed

c) No description of the number and reason for refusing assessment.

4) Ascertainment of the assessment (risk factor):

a) Validated measurement tool. **

b) Non-validated measurement tool, but the tool is available or described. *

c) No description of the measurement tool.

Comparability: (Maximum 2 stars)

1) The subjects in different outcome groups are comparable, based on the study design or analysis. Confounding factors are controlled.

a) The study controls for the most important factor (select one). *

b) The study control for any additional factor. *

Outcome: (Maximum 3 stars)

1) Assessment of the outcome:

a) Independent blind assessment. **

b) Record linkage. **

c) Self report. *

d) No description.

2) Statistical test:

a) The statistical test used to analyse the data is clearly described and appropriate, and the measurement of the association is presented, including confidence intervals and the probability level (p value). *

b) The statistical test is not appropriate, not described or incomplete.

APPENDIX S5: Updated criteria for Good Measurement Properties

Measurement property	Rating	Criteria
Reliability	+	ICC or weighted Kappa ≥ 0.70
	?	ICC or weighted Kappa not reported
	-	ICC or weighted Kappa < 0.70
Measurement error	+	Smallest detectable change (SDC) or limits of agreement (LoA) < minimal important change (MIC)
	?	MIC not defined
	-	SDC or LoA > MIC
Hypotheses testing for construct validity	+	The result is in accordance with the hypothesis
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis
Cross-cultural validity/ measurement invariance	+	No important differences found between group factors (such as age, gender, language) in multiple group factor analysis OR no important DIF for group factors (McFadden’s R < 0.02)
	?	No multiple group factor analysis OR DIF analysis performed
	-	Important differences between group factors OR DIF was found
Criterion validity	+	Correlation with gold standard ≥ 0.70 OR AUC ≥ 0.70
	?	Not all information for ‘+’ reported
	-	Correlation with gold standard < 0.70 OR AUC < 0.70
Responsiveness	+	The result is in accordance with the hypothesis OR AUC ≥ 0.70
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis OR AUC < 0.70

Adapted from Prinsen CA, et al.[2] then modified by removing structural validity and internal consistency item.

APPENDIX S6: Modified GRADE approach for grading the quality of evidence

Quality of evidence	Lower if
High	Risk of bias
Moderate	−1 Serious
Low	−2 Very serious
Very low	−3 Extremely serious
	Inconsistency
	−1 Serious
	−2 Very serious
	Imprecision
	−1 total n = 50–100
	−2 total n < 50
	Indirectness
	−1 Serious
	−2 Very serious

The starting point is the assumption that the evidence is of high quality. The quality of evidence is subsequently downgraded with one or two levels for each factor (i.e., risk of bias, inconsistency, imprecision, indirectness) to moderate, low, or very low when there is risk of bias (low study quality), (unexplained) inconsistency in results, or indirect results.[3] Information on how to downgrade is described in detail in the COSMIN user manual.[1] n = sample size.

Appendix S7. Definition of quality levels

Quality Level	Definition
High	We are very confident that the true measurement property lies close to that of the estimate of the measurement property
Moderate	We are moderately confident in the measurement property estimate: the true measurement property is likely to be close to the estimate of the measurement property, but there is a possibility that it is substantially different
Low	Our confidence in the measurement property estimate is limited: the true measurement property may be substantially different from the estimate of the measurement property
Very low	We have very little confidence in the measurement property estimate: the true measurement property is likely to be substantially different from the estimate of the measurement property

These definitions were adapted from the GRADE approach.[4] Information on how to downgrade is described in detail in the COSMIN user manual.[1]

REFERENCES

1. Mokkink LB, Prinsen C, Patrick DL, et al. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User manual* 2018. https://cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf
2. Prinsen CA, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018; **27**: 1147-57.
3. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol*. 2011; **64**: 383-94.
4. Schünemann H, Brozek J, Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations. Updated October 2013. https://training.cochrane.org/resource/grade-handbook

SUPPLEMENTARY DATA

APPENDIX S1: Search strategy

Search strategy for Ovid Medline Version 15/08/20

PICO

Population

Postoperative patients aged 18 years and over from all surgical disciplines.

Intervention

Unidimensional pain assessment tools including

- ◊ Verbal or printed numerical pain rating scale.
- ◊ Printed or verbal descriptor scale.
- ◊ Visual analogue scale.
- ◊ Faces scales: Wong-baker FACES, Faces Pain Scale – Revised.

Functional pain assessment tools

Comparison: -------

Outcomes: psychometric properties including validity and reliability

Additional outcomes

Instrument feasibility, interpretability, and ability to detect desire of analgesia.

Search concepts to be combined for Boolean AND, and used for unidimensional pain assessment tool and then repeated for functional pain assessment tools

1. Outcome terms
2. Pain assessment tool terms

3. Construct: acute postoperative pain
4. 1 AND 2 AND 3
5. 4 + Limits (english , humans, adults > 18 years)

Did not apply limits full text, abstracts this might include bias in the results

Ovid MEDLINE(R) ALL < 1946 to August 15, 2022>

1 exp PSYCHOMETRICS/ or psychometr*.mp. or measurement propert*.mp. or Validity.mp. or valid*.mp. or exp Validation Study/ or convergent validity.mp. or construct validity.mp. or content validity.mp. or criterion validity.mp. or reliab*.mp. or unreliab*.mp. or Comparative Study.mp. or Feasibility.mp. or Generalizability.mp. or generalisa*.mp. or interpretab*.mp. or Sensitiv*.mp. or Responsive*.mp. or 'Measurement Accuracy'.mp. or 'ease of use'.mp. or Analgesi* response.mp. or 'desire of analgesi*'.mp. or 'Request of analgesic*'.mp. or 'hypotheses testing'.mp. or 'measurement error*'.mp. or Internal consistency.mp. or Data accuracy.mp. or 'standard error of measurement'.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 4890505

2 (pain scale* or pain rating scale* or (pain assessment and (instrument* or tool*)) or pain intensity scale* or pain measurement instrument* or Pain score* or pain intensity assessment).mp. or exp Pain Measurement/ [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 113996

3 Visual Analog Scale.mp. or exp Visual analog? Pain scale/ or (visual analog? and (scale or score)).mp. or vas.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 146135

4 ((numeric* and rating and (scale or score)) or numeric scale or nrs or nprs).mp. or exp numerical pain rating scale/ [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 26611

5 exp verbal descriptor scale/ or Vds.mp. or exp verbal rating scale*/ [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 1128

6 exp face* pain scale*/ or exp wong baker Face*/ or wong baker face*.mp. or exp faces pain scale revised/ or faces pain scale revised.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 594

7 (pain activity assessment or functional pain assessment scale or functional activity score*or functional pain activity scale* or functional assessment tool or objective pain score* or movement evoked pain assessment or assessment of pain at movement or objective pain assessment or clinically aligned pain assessment tool).mp. [mp=title, abstract,

original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 252

8 exp Pain, Postoperative/ or exp acute pain/ or post surgical pain.mp. or surgical pain.mp. or pain post procedure.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 46322

9 1 and 3 and 8 5987

10 1 and 4 and 8 556

11 1 and 5 and 8 6

12 1 and 6 and 8 56

13 1 and 7 and 8 32

14 limit 9 to (Elanguage and humans and "all adult (19 plus years)") 4358

15 limit 10 to (English language and humans and "all adult (19 plus years)") 537

16 limit 11 to (English language and humans and "all adult (19 plus years)") 6

17 limit 12 to (English language and humans and "all adult (19 plus years)") 12

18 limit 13 to (English language and humans and "all adult (19 plus years)") 2

Search strategy for other databases can be provided on demand from the corresponding author

APPENDIX S2: Measurment properties included in the main domians of the COSMIN taxonomy

Domain	Psychometric property	Definition
Reliability		The extent that the measurement is free from measurement error such that scores for patients who have not changed are the same under repeated measurements
	Internal consistency	The extent that items are inter-related
	Reliability	The proportion of the total variance in the measurements that is due to 'true' differences between patients (as opposed to error)
	Measurement error	Error in a participant's score that is not attributed to the construct being measured
Validity		The extent that an assessment measures what it aims to measure
	Content validity	The extent that an assessment's content reflects the construct being measured
	Face validity	The extent that an assessment looks like it reflects the construct being measured
	Construct validity	The extent that an assessment's scores are consistent with hypotheses based on the assumption that the tool measures what it purports to measure
	Structural validity	The extent that an assessment's scores reflect the dimensionality of the construct being measured
	Hypothesis testing	
	Cross-cultural validity	Construct validity for the items of an assessment
		The extent that items on a translated or culturally modified assessment reflect the original items
	Criterion validity	The extent that an assessment's scores represent the 'gold standard'

Responsiveness	An assessment and/or it’s items’ ability to detect change over time in the construct being measured
Interpretability*	The extent that clinical or everyday understanding can be applied to an assessment’s scores
Feasibility*	How easily a pain measure can be scored and interpreted

COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments. Adopted from Mokkink LB, et al.[1]

*Interpretability and *feasibility are not considered measurement properties, but important characteristics of a measurement instrument.

APPENDIX S3: Studies ineligible following full-text review

Full paper examined: 38/ Exclusion after complete paper screening 19 papers.

Excluded papers:

1. Arnstein P, Gentile D, Wilson M. validating the functional pain scale for hospitalized adults. *Pain Manag Nurs*. 2019; **20:** 418-24.

Explanation: Paper validating functional scale for hospitalized chronic pain patient but did not report separte result for surgical patients.

Reason for exclusion: No separate results for postoperative pain assessment.

2. Barber MD, Janz N, Kenton K, et al. Validation of the surgical pain scales in women undergoing pelvic reconstructive surgery. *Female Pelvic Med Reconstr Surg*. 2012; **18:** 198-204.

Explanation: Surgical pain scale looked at long term functional outcome following surgery.

Reason for exclusion: Patients not assessed as inpatients/irrelevant outcome.

3. McCarthy Jr M, Chang CH, Pickard AS, et al. Visual analog scales for assessing surgical pain. *Jl Amn Coll Surg*. 2005; **201:** 245-52.

Reason for exclusion: Patients not assessed as inpatients or irrelevant outcome.

4. Blumstein HA, Moore D. Visual analog pain scores do not define desire for analgesia in patients with acute pain. *Acad Emerg Med.* 2003; **10:** 211-4.

Explanation: VAS to detect desire of analgesia in acute emergency pain.

Reason for exclusion: Not surgical population.

5. Chiu LYL, Sun T, Ree R, et al. The evaluation of smartphone versions of the visual analogue scale and numeric rating scale as postoperative pain assessment tools: a prospective randomized trial. *Can J Anesth*. 2019; **66:** 706-15.

Reason for exclusion: Comparison between NRS smart version with paper version.

6. Neudecker J, Raue W, Schwenk W. High correlation but inadequate point-to-point agreement, between conventional mechanical and electronical visual analogue scale for assessment of acute postoperative pain after general surgery. *Acute Pain*. 2006; **8:** 175-80.

Reason for exclusion: Comparison between electronic and mechanical VAS.

7. Erden S, Karadag M, Guler Demir S, et al. Cross-cultural adaptation, validity, and reliability of the Turkish version of revised American Pain Society patient outcome questionnaire for surgical patients. *Agri*. 2018; **30:** 39-50.

Reason for exclusion: Multidimensional tool (Revised American Pain Society Patient Outcome Questionnaire).

8. Keawnantawat P, Thanasilp S, Preechawong S. Translation and validation of the Thai version of a modified brief pain inventory: a concise instrument for pain assessment in postoperative cardiac surgery. *Pain Pract*. 2017; **17:** 763-73.

Reason for exclusion: Multidimensional tool (modified brief pain inventory).

9. Mendoza TR, Chen C, Brugger A, et al. The utility and validity of the modified Brief Pain Inventory in a multiple-dose postoperative analgesic trial. *Clin J Pain*. 2004; **20:** 357-62.

Reason for exclusion: Multidimensional tool (Brief Pain Inventory).

10. Mwachiro M, Mwachiro E, Wachu M, et al. assessing post-operative pain with self-reports via the Jerrycan Pain Scale in Rural Kenya. *World J Surg*. 2020; **44:** 3636-42.

Reason for exclusion: Applicability of irrelevant tool (Jerrycan Pain Scale).

11. Jain R, Grewal A. A randomized comparative study assessing efficacy of pain versus comfort scores. *Saudi J Anaesth*. 2017; **11**: 396-401.

Reason for exclusion: Retracted paper.

12. Liu WH, Aitkenhead AR. Comparison of contemporaneous and retrospective assessment of postoperative pain using the visual analogue scale. *Br J Anaesth*. 1991; **67**: 768-71.

Reason for exclusion: Irrelevant outcome.

13. Salo D, Eget D, Lavery RF, Garner L, Bernstein S, on K. Can patients accurately read a visual analog pain scale? *Am J Emerg Med*. 2003; **21**: 515-9.

Reason of exclusion: Not surgical population.

14. Sills ES, Genton MG, Walsh APH, Wehbe SA. Who's asking? Patients may under-report postoperative pain scores to nurses (or over-report to surgeons) following surgery of the female reproductive tract. *Arch Gynecol Obstet*. 2009; **279**: 771-4.

Explanation: Looked at how patient communicate pain between nurse and physician.

Reason for exclusion: Irrelevant outcome.

15. Rothaug J, Weiss T, Meissner W. How simple can it get? Measuring pain with NRS items or binary items. *Clin J Pain*. 2013; **29**: 224-32.

Explanation: They used different answer format for (binary yes/no answers vs. NRS) in a subset of patients using Quality Improvement in Postoperative Pain Management (QUIPS).

Reason for exclusion: Multidimensional tool (QUIPS).

16. Zalon ML. Comparison of pain measures in surgical patients. *J Nurs Meas*. 1999; **7:** 135-52.

Explanation: This study aimed to establish the validity of brief pain inventory short form.

Reason for exclusion: Validation of multidimensional scale.

17. Halm M, Bailey C, St Pierre J, et al. Pilot evaluation of a functional pain assessment scale. *Clin Nurse Spec*. 2019; **33:** 12-21.

Explanation: Sample from medical/surgical, critical care, and rehabilitation units experiencing acute or chronic pain.

Reason for exclusion: No separate results for acute postoperative pain.

18. Martin WJJM, Ashton-James CE, Skorpil NE, et al. What constitutes a clinically important pain reduction in patients after third molar surgery? *Pain Res Manag*. 2013; **18:** 319-22.

Reason for exclusion: Dental surgery, not hospitalized patients.

19. Rago R, Forfori F, Materazzi G, et al. Evaluation of a preoperative pain score in response to pressure as a marker of postoperative pain and drugs consumption in surgical thyroidectomy. *Clin J Pain*. 2012; **28:** 382-6.

Reason for exclusion: Sensitivity of preoperative vas scores after tourniquet pressure inflation.

APPENDIX S4: Newcastle-Ottawa Quality Assessment Scale

(adapted for cross sectional studies)

This scale has been adapted from the Newcastle-Ottawa Quality Assessment Scale for cohort studies to perform a quality assessment of cross-sectional studies for the systematic review.

Selection: (Maximum 4 stars)

1) Representativeness of the sample:

a) Truly representative of the average in the target population. * (all subjects or random sampling)

b) Somewhat representative of the average in the target population. * (non-random sampling)

c) Selected group of users.

d) No description of the sampling strategy.

2) Sample size:

a) Justified and satisfactory. (by reporting appropriate sample size calculation) *

b) Not justified.

3) Non-respondents: (adopted to details about patient refused assessment and reasons are described)

a) Comparability between assessed and non-assessed is established *

b) The response rate is unsatisfactory, or the comparability between respondents and non-respondents is unsatisfactory. removed

c) No description of the number and reason for refusing assessment.

4) Ascertainment of the assessment (risk factor):

a) Validated measurement tool. **

b) Non-validated measurement tool, but the tool is available or described. *

c) No description of the measurement tool.

Comparability: (Maximum 2 stars)

1) The subjects in different outcome groups are comparable, based on the study design or analysis. Confounding factors are controlled.

a) The study controls for the most important factor (select one). *

b) The study control for any additional factor. *

Outcome: (Maximum 3 stars)

1) Assessment of the outcome:

a) Independent blind assessment. **

b) Record linkage. **

c) Self report. *

d) No description.

2) Statistical test:

a) The statistical test used to analyse the data is clearly described and appropriate, and the measurement of the association is presented, including confidence intervals and the probability level (p value). *

b) The statistical test is not appropriate, not described or incomplete.

APPENDIX S5: Updated criteria for Good Measurement Properties

Measurement property	Rating	Criteria
Reliability	+	ICC or weighted Kappa ≥ 0.70
	?	ICC or weighted Kappa not reported
	-	ICC or weighted Kappa < 0.70
Measurement error	+	Smallest detectable change (SDC) or limits of agreement (LoA) < minimal important change (MIC)
	?	MIC not defined
	-	SDC or LoA > MIC
Hypotheses testing for construct validity	+	The result is in accordance with the hypothesis
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis
Cross-cultural validity/ measurement invariance	+	No important differences found between group factors (such as age, gender, language) in multiple group factor analysis OR no important DIF for group factors (McFadden's R < 0.02)
	?	No multiple group factor analysis OR DIF analysis performed
	-	Important differences between group factors OR DIF was found
Criterion validity	+	Correlation with gold standard ≥ 0.70 OR AUC ≥ 0.70
	?	Not all information for '+' reported
	-	Correlation with gold standard < 0.70 OR AUC < 0.70
Responsiveness	+	The result is in accordance with the hypothesis OR AUC ≥ 0.70
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis OR AUC < 0.70

Adapted from Prinsen CA, et al.[2] then modified by removing structural validity and internal consistency item.

APPENDIX S6: Modified GRADE approach for grading the quality of evidence

Quality of evidence	Lower if
High	Risk of bias
Moderate	−1 Serious
Low	−2 Very serious
Very low	−3 Extremely serious
	Inconsistency
	−1 Serious
	−2 Very serious
	Imprecision
	−1 total n = 50–100
	−2 total n < 50
	Indirectness
	−1 Serious
	−2 Very serious

The starting point is the assumption that the evidence is of high quality. The quality of evidence is subsequently downgraded with one or two levels for each factor (i.e., risk of bias, inconsistency, imprecision, indirectness) to moderate, low, or very low when there is risk of bias (low study quality), (unexplained) inconsistency in results, or indirect results.[3] Information on how to downgrade is described in detail in the COSMIN user manual.[1] n = sample size.

Appendix S7. Definition of quality levels

Quality Level	Definition
High	We are very confident that the true measurement property lies close to that of the estimate of the measurement property
Moderate	We are moderately confident in the measurement property estimate: the true measurement property is likely to be close to the estimate of the measurement property, but there is a possibility that it is substantially different
Low	Our confidence in the measurement property estimate is limited: the true measurement property may be substantially different from the estimate of the measurement property
Very low	We have very little confidence in the measurement property estimate: the true measurement property is likely to be substantially different from the estimate of the measurement property

These definitions were adapted from the GRADE approach.[4] Information on how to downgrade is described in detail in the COSMIN user manual.[1]

REFERENCES

1. Mokkink LB, Prinsen C, Patrick DL, et al. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User manual* 2018. https://cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf
2. Prinsen CA, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018; **27**: 1147-57.
3. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol*. 2011; **64**: 383-94.
4. Schünemann H, Brozek J, Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations. Updated October 2013. https://training.cochrane.org/resource/grade-handbook