PAIN Publish Ahead of Print
DOI: 10.1097/j.pain.0000000000000689

Title: Reliability of Conditioned Pain Modulation: a Systematic Review

Authors: Donna L. Kennedy[1], Harriet I. Kemp[1], Deborah Ridout[2], David Yarnitsky[3], Andrew SC Rice[1]

Author Affiliations: 1. Pain Research, Imperial College London, London, United Kingdom

2. Institute of Child Health, University College London, London, United Kingdom

3. Neurology, Rambam Health Care Campus, Technion Faculty of Medicine, Haifa, Israel

Number of text pages:

Author for correspondence:
Donna L. Kennedy
Pain Research Group
Imperial College London
Chelsea & Westminster Campus
369 Fulham Road
London SW10 9NH
United Kingdom
+44 (0)208 746 8424
Email: d.kennedy@imperial.ac.uk

Number of pages in manuscript – 22
Number of figures – 2
Number of tables - 3

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

Abstract: A systematic literature review was undertaken to determine if conditioned pain modulation (CPM) is reliable. Longitudinal, English language observational studies of the repeatability of a CPM test paradigm in adult humans were included. Two independent reviewers assessed the risk of bias in six domains; study participation; study attrition; prognostic factor measurement; outcome measurement; confounding and analysis using the Quality in Prognosis Studies (QUIPS) critical assessment tool [17]. Intraclass correlation coefficients (ICCs) less than 0.4 were considered to be poor; 0.4 and 0.59 to be fair; 0.6 and 0.75 good and greater than 0.75 excellent [37]. Ten studies were included in the final review. Meta-analysis was not appropriate due to differences between studies. The intersession reliability of the CPM effect was investigated in 8 studies and reported as good (ICC = 0.6-.75) in 3 studies and excellent (ICC > .75) in subgroups in 2 of those 3. The assessment of risk of bias demonstrated that reporting is not comprehensive for the description of sample demographics, recruitment strategy and study attrition. The absence of blinding, a lack of control for confounding factors and lack of standardisation in statistical analysis are common.   CPM is a reliable measure, however the degree of reliability is heavily dependent upon stimulation parameters and study methodology and this warrants consideration for investigators. The validation of CPM as a robust prognostic factor in experimental and clinical pain studies may be facilitated by improvements in the reporting of CPM reliability studies.

Keywords: Conditioned pain modulation (CPM); diffuse noxious inhibitory control (DNIC); endogenous pain modulation; reliability; systematic review

BACKGROUND

Conditioned Pain Modulation

Conditioned pain modulation (CPM) is a psychophysical experimental measure of the endogenous pain inhibitory pathway in humans; the "pain inhibits pain" phenomena [41]. CPM is believed to represent the human behavioural correlate of diffuse noxious inhibitory control (DNIC), first described in rats [22]. Electrophysiological studies in animals and pharmacological studies in humans have demonstrated that descending influences on spinal nociceptive processing involve the periaqueductal gray (PAG) rostral ventromedial medulla (RVM) and subnucleus reticularis dorsalis, leading to the description of this descending pain modulation pathway as a spino-bulbo-spinal loop [27].

CPM paradigms consist of the evaluation of a painful test stimulus followed by a second evaluation either at the same time as a distant, painful conditioning stimulus (parallel paradigm) or in series after the painful conditioning stimulus has been withdrawn (sequential paradigm) [41]. While pain inhibition is not universal; in some subjects an increase in pain intensity rating is observed (facilitation), in the majority of subjects the pain intensity experienced with the test stimulus will be reduced during or immediately following exposure to the conditioning stimulus.

CPM has been investigated extensively in healthy volunteers, however at present there are no published normative data for CPM effect and it is unclear what qualifies as a "normal range" effect. In a review of healthy volunteer studies Pud et al. [35] reported variability in the magnitude of CPM effect was dependent upon the CPM paradigm employed and that the median CPM effect was 29%. However, this must be interpreted with some caution given the heterogeneity and lack of quality assessment of the included studies. There is good evidence that there is much inter-individual difference in the magnitude of CPM related to age, sex and potentially other as yet unknown variables [9; 10]. It has been reported that in some healthy subjects a CPM effect may be altogether absent [25], although it is probably more accurate to consider that the spectrum of response may range from significant inhibition to a degree of facilitation dependent upon individual variability and CPM paradigm. In healthy volunteer studies, the appreciable variability reported in magnitude and the stability of the CPM effect may be attributable to multiple factors including variation in study characteristics such as study design and testing parameters or variability in sample characteristics as defined by the inclusion and exclusion criteria used to qualify a sample of volunteers as "healthy" [7; 11].

At present there is great interest in the science and conduct of CPM testing as there is a growing body of evidence suggesting that CPM may be an important biomarker of chronic pain as well as a predictor of treatment response. However standardization in the testing of CPM is lacking. A 2014 consensus meeting encouraged investigators to include a second test stimulus or second CPM protocol in study designs for the generation of evidence to enable comparisons, suggested sequential test protocols may be advantageous over parallel protocols for being a purer measure of CPM, and that an upper and lower limb should be default test sites, however the expert forum concluded that there was insufficient data to support recommendations for the use of a specific CPM protocol [42] and this has not changed to date. There is evidence to suggest that the magnitude of the CPM effect is dependent upon the sensory modality employed for delivering the conditioning and test stimuli and the body area tested [29; 35] as well as the painfulness of the stimuli [13] however at present there is no gold standard for the testing of CPM. Furthermore, estimating the reliability of CPM, as well as identifying true change in relation to measurement error, has proven challenging due to heterogeneity in study design and analysis and insufficient reporting.

OBJECTIVES

To assess the reliability of CPM paradigms in adults, critically appraise the literature against reporting guidelines for prognostic factor research and CPM studies [41; 42] and make recommendations for the reporting of future studies.

METHODS

The protocol for this review was not registered as it does not meet the inclusion criteria of the available web-based repositories. Findings are reported according to the PRISMA guidelines for systematic reviews [28].

Literature Search

No previously published systematic reviews of the reliability of CPM were located neither in the Cochrane Database of Systematic Reviews nor in a search of the electronic databases MEDLINE, EMBASE, CINAHL and AMED. The same databases were searched from inception to August 26th 2015 using the search terms (conditioned pain modulation or diffuse noxious inhibitory control or DNIC or heterotopic noxious conditioning) and (reliability or repeatability or stability) (Appendix A). Inclusion criteria were full- text English reports of longitudinal observational studies of the repeatability or stability of a CPM test paradigm in adult humans. Two independent reviewers (D.K., H.K.) screened study titles, abstracts and where necessary full-text to determine study inclusion (Figure 1). Reference lists of included studies were hand searched for additional eligible studies.

Data Extraction and Management

Two review authors independently extracted data using a standardized form (D.K., H.K.). This included sample size, participant gender and mean age, designation as a healthy volunteer or clinical cohort, test and conditioning stimuli and testing site, testing paradigm (sequential or parallel), re-test interval, reliability coefficient for CPM effect, measure of response stability, protocol violations (any deviation from a study protocol that may affect the reliability of the data) and test and conditioning stimulus reliability.

Risk of Bias Assessment

The methodological quality and risk of bias of the included studies was assessed by two independent raters (D.K., H.K.) using the Quality in Prognosis Studies (QUIPS) critical assessment tool; a tool specifically developed for use in systematic reviews of prognostic factor studies [16]. The QUIPS appraisal domains are in keeping with the National Institutes for Health (NIH) mandate to improve rigor, transparency and reproducibility in research [8; 21]. For clarity, while published CPM reliability studies do not purport to be prognostic factor studies, it is our intent to initiate and encourage future work toward strengthening the evidence for CPM as a prognostic factor. The QUIPS tool addresses risk of bias in six major domains; study participation; study attrition; prognostic factor measurement; outcome measurement; confounding and statistical analysis and is designed to be operationalized for specific study purposes including specifying key characteristics, omitting irrelevant items and adding items where required [17]. Criteria in each domain are evaluated, thereby generating an overall rating for each domain as having a" low", "moderate" or "high" risk of bias. For this review, the QUIPS tool was operationalized to be study specific a priori and is reported in Appendix C. This descriptive approach to quality assessment in systematic reviews is in keeping with current recommendations given the questionable validity and interpretation of existing rating scales [18].

Appraisal of Reliability Data

Reliability data was included in the risk of bias in statistical analysis and interpreted as a measure of the repeatability of a CPM paradigm. Important elements in the statistical analysis of reliability include the reporting of a sample size calculation, an appropriate reliability coefficient and 95% confidence interval for the coefficient and a measure of response stability. Where any of these components were lacking this was interpreted to increase the risk of bias in statistical analysis and reporting.

While there is lack of consensus in the appropriate analysis and reporting of reliability for measures which produce continuous data, as does CPM, there is growing evidence to support the use of the intraclass correlation coefficient (ICC) which reflects both the degree of association and agreement among ratings [34; 36; 37]. Because the ICC is a dimensionless statistic, it is also useful when comparing the repeatability of measures in different units [5]. There are three models of ICCs; the choice of model is fundamental in assessing the reliability of clinical or experimental tests and must consider if the use of an instrument or procedure may be generalised to a wider population of random raters, or if performance is user-dependent, perhaps reflecting specialist training.

The ICC has been described as a measure of relative reliability as it reflects the degree to which a subject maintains their place in a sample [1], however reported in isolation the ICC gives no indication of the magnitude of the disagreement between measures or retests [36]. Response stability, also described as absolute reliability [1] describes the degree to which a subject's scores will change over repeated tests. A measure of response stability is essential to the practical and clinical interpretation of reliability. While the ICC provides a dimensionless and easily interpreted point estimate of reliability, a measure of response stability facilitates the comparison of results between reliability studies and enables the judgement of when a change in test score is clinically meaningful rather than due to measurement error. While reliability cannot be interpreted as an all or none concept and acceptable reliability is subjective, there is some consensus that a coefficient less than 0.4 may be interpreted as poor reliability; between 0.4 and 0.59 fair reliability; between 0.6 and 0.75 good reliability; and greater than 0.75 excellent reliability therefore the reliability coefficients reported in this review were interpreted as such [37].

(Insert figure 1 here)

RESULTS

Ten studies were selected for inclusion in this review (see Fig. 1). At screening, excluded records did not pertain to the reliability of CPM or were not full text papers. One full-text article was excluded and is reported in Appendix B. No full text papers examining the reliability of a CPM paradigm were excluded.

Study Characteristics

Summary information for the included studies is reported in Table 1. Seven studies investigated CPM in healthy volunteers, two studies addressed clinical cohorts and one study included both healthy subjects and a clinical cohort. Eight studies included males and females; one study had only males, one only females. In healthy subject studies the participants were predominantly under the age of 40 while clinical cohort participants were predominantly over the age of 40.

The most commonly investigated test stimulus was pressure pain threshold (5 studies), followed by contact heat pain (3 studies). Cold water immersion was the most frequently studied conditioning stimulus (6 studies) followed by hot water immersion (3 studies) and Ischemic pain (3 studies). Inter-session reliability was investigated in 9 studies with re-test intervals varying between 2 and 28 days; intra-session reliability was investigated in 3 studies. The most commonly reported outcome measures were subjective pain threshold (6 studies) and an individualised stimulus intensity required to elicit a pre-determined pain intensity (5 studies). Subjective pain intensity rating was measured in 2 studies, a pain elicited reflex in 2 studies and subjective pain tolerance in 1 study.

Where reported, study protocol violations and the reliability coefficient for the test and conditioning stimuli are reported in Table 2. Protocol violations for the administration of the test and conditioning stimulus include changes to exposure time or intensity of the stimulus from that described a priori and in which case the participant was not excluded from the study. There were no reported study violations in the administration of the test stimuli and 3 reported protocol violations for cold water immersion as a conditioning stimulus.

Reliability of CPM Effect

The intra-session reliability of the CPM effect was investigated in 9 different test-retest measures in 3 studies and was reported as good (ICC = .6-.75) to excellent (ICC >75) in 7 of 9 measures. The intersession reliability of the CPM effect was investigated in 14 different testing paradigms (different test stimuli, outcome measures, pain intensity) in 8 studies. Investigators in 6 out of 8 studies reported intersession reliability ranging from fair to excellent for a CPM paradigm. Poor intersession reliability was reported for the CPM effect in older adults with chronic pancreatitis and in young women across menstrual cycles (Table 1).

Reliability of Test Stimuli

Pressure pain threshold was most commonly employed as a test stimulus; intra-session reliability was reported as excellent in 2 studies (ICC > .75); intersession reliability as good in 2 studies (ICC =.60-.75) and excellent in 1 study. The reliability of contact heat pain was reported in two studies. Where a thresholding technique was used to individualise the temperature required to elicit pain at a pre-determined intensity, the repeatability of the test stimulus temperature ranged from fair to excellent (ICC = .53; ICC =.64; ICC=.83). In contrast, the subjective pain rating for the contact heat pain test stimulus ranged from poor to fair (ICC =. 19; ICC =. 31; ICC =.4). The reliability of a pain elicited reflex was reported in 2 studies and ranged from good to excellent (ICC .61; ICC = .93) (Table 2).

Reliability of Conditioning Stimuli

Five studies investigated the intersession reliability of a conditioning stimulus by comparing subjective pain ratings for the stimulus from 2 test sessions. The reliability of pain ratings for immersion in a hot water bath range from fair to excellent (ICC=. 54; ICC = .76; ICC= .79); for immersion in cold water good to excellent (ICC = .61; ICC = .80) and for ischemic pain excellent (ICC = .82). Poor reliability (ICC = .16) was reported for contact heat pain ($Pain^{30}$+ .5°C) as a conditioning stimulus (Table 2).

Risk of Bias in Included Studies

Results for the assessment of risk of bias are reported in Table 3. A moderate to high risk of bias for study participation and study attrition was found. The risk of bias for prognostic factor measurement was moderate as reporting of investigator or participant blinding was lacking. Risk of bias in study confounding ranged from low to high; for outcome measurement was assessed as low and for risk of bias in statistical analysis and reporting was moderate to high.

(Insert Table 1 here)

(Insert Table 2 here)

(Insert Table 3 here)

DISCUSSION

Summary of Results

The aim of this review was to determine if CPM is reliable. This review incorporated 9 studies reporting 23 test-retest measures of various CPM test paradigms in heterogeneous populations and therefore meta-analysis of results was not appropriate. However, 78% of reported reliability coefficients for the intra-session reliability were interpreted as good (ICC =.6 - .75) or excellent (ICC> .75). Intersession reliability was reported in 8 studies and reliability coefficients were interpreted as good or excellent in 50% of studies. The reliability of a CPM paradigm is dependent on test and conditioning stimulus, stimulation parameters, test sites and study population.

Reporting and Risk of Bias (Table 3)

In this review, there was a moderate to high risk of bias for both study participation and study attrition. A recently published consensus paper defines the characteristics of healthy subjects in quantitative sensory testing studies [11]. In order for the reader to ascertain susceptibility to bias, we suggest in future studies the source of the target population, the sampling frame and methods of recruitment, the place or places and dates of recruitment, study inclusion and exclusion criteria, the numbers recruited to the numbers enrolled and baseline characteristics of the study sample be reported. In addition to facilitating the assessment of risk of bias, more thorough description of a study sample aides the generalization of results to other populations.

The aim in rating risk of attrition bias is determining the possibility that the prognostic factor, in this case CPM effect, is different for those who complete versus those who do not complete the study. Generally a moderate risk of attrition bias was found. Study drop-outs were not consistently reported, nor was information provided on key characteristics of those who dropped out of the studies which would have enabled an appraisal of whether those who dropped out differed systematically from those who continued in the study.

The risk of bias for prognostic factor measurement was generally moderate; reporting of investigator or participant blinding was lacking. While assessor blinding is challenging in measures such as CPM, future investigations might consider how this can be addressed. For the majority of studies, it is unclear what information the participants received regarding the experiment which may have influenced their response or created expectation, or what their exposure was between intersession measures. Additionally, there was lack of detail regarding the standardization of test instructions between participants and in a number of studies the conditioning stimulus was not consistent for all participants.

Risk of bias in study confounding ranged from low to high. In healthy volunteer studies, common exclusions included pain conditions, pain medication and psychiatric history. However, it was common that baseline and retest measures of health and pain were not employed, making the assumption that participants were indeed pain free at retest. While it is difficult to interpret the effect of confounding on reliabilty, it would appear there may be an association. In studies of intersession reliability, there appears to be a trend, with lower risk of bias in confounding associated with greater reliability. This would suggest that in studies with lower risk of bias, important factors that may influence the CPM effect were controlled for between sessions, thereby improving repeatability.

(Insert Figure 2 here)

The risk of bias in statistical analysis and reporting was rated as moderate to high. The publication dates of the studies included in this review range from 2009 to 2015 and while the reporting of statistical methods has improved with subsequent publications, it is important that improvements continue to be made in this area. As noted previously, the precision of a reliability coefficient is dependent of an appropriate sample size and at present sample size calculations are generally lacking in CPM reliability studies. And while the model of ICC used for statistical analysis should be reported, this has been consistently under-reported.

It is clear that reducing risk of bias in the conduct and reporting of CPM reliability studies is essential to improve transparency and make gains towards the identification of robust, reliable CPM paradigms. At present, a moderate to high risk of bias for prognostic factor measurement may be introducing random error into testing, and thereby reducing reliability. As noted above, the same may be said for risk of bias in confounding, with lack of control for important participant- related variables subsequently reducing re-test reliability. In contrast, risk of bias for study participation, study attrition and analysis and reporting may be unintentionally over- inflating reliability estimates. It is only with improved rigour in study design and reporting that we can move toward standardisation in testing.

Reliability of Test and Conditioning Stimulus (Table 2)

While the test and conditioning stimulus must be noxious, the methods and parameters for delivering these stimuli vary. If a test or conditioning stimulus is overly painful, it is possible that it may not be tolerated by all participants and therefore the stimulus is not applied uniformly to the sample. There is evidence to suggest that the repeatability of the various test and conditioning stimuli vary across sessions, and this lack of repeatability of the components of the CPM paradigm may reduce the repeatability for the sum total of the paradigm.

For the studies included in this review, there were no reports of participants not tolerating the test stimulus (PPT, contact heat, nociceptive withdrawal or flexion reflexes) as specified in the study protocols, therefore creating a protocol violation. As the test stimuli described are phasic, this brief exposure to a noxious stimulus appears well tolerated. In comparison, the conditioning stimuli reported (ischemic pain, cold pressor test, contact heat, hot water bath, contact heat) are tonic, vary in intensity and exposure and in how well they are tolerated by participants. Using ischaemic pain [33] and contact heat [14] as conditioning stimuli, there were no reported participant withdrawals, i.e. all participants tolerated the stimulus for the time period specified in the protocol. In contrast, participant tolerance to immersion in the cold pressor test (CPT) and hot water bath appear time and temperature dependent. This suggests that CPT temperatures of between 8° and 12°C and for up to 2 minutes and hot water bath immersion at 46.5°C for 1 minute are sufficient to induce inhibition and are well tolerated by participants, ensuring that the conditioning stimulus is consistent for all participants and thereby perhaps improving repeatability. This is consistent with the findings of Granot et al. [13] regarding the intensity of heat and cold pain necessary to induce CPM. These findings have important implications for the investigation of CPM paradigms in populations with chronic, painful conditions; if a stimulus is not well tolerated by a sample of healthy volunteers it is perhaps even less likely to be tolerated by patients who are in pain.

The Reliability of Parallel versus Sequential Paradigms (Table 1)

Two studies, Olesen et al. [32] and Valencia et al. [39] investigated sequential CPM paradigms with reliability reported as poor, and good to excellent, respectively. The remainder investigated parallel paradigms with intersession reliability ranging from poor to good therefore it is impossible to conclude from the available evidence if there is greater reliability for one paradigm over another.

Timing of intra-session assessments (Table 1)

For the three studies that investigated intrasession reliability, the wash-out period between intrasession assessments included 2 minutes; 15 minutes and 60 minutes [6; 23; 38], respectively. With a 2 minute wash-out reliability ranged from fair to good, for 15 minutes good to excellent, and for 60 minutes fair to good therefore it is difficult to discern the impact of wash-out time on intra-session reliability from this review.

Non-Responders

An important consideration in the clinical or experimental utility of a CPM paradigm is whether or not the paradigm induces a CPM effect and, if so, in what proportion of subjects. While the reporting of absolute and percentage change in CPM effect speaks to the magnitude of change, that is, the reduction in pain ratings or increase in threshold of the test stimulus following exposure to the conditioning stimulus, this approach does not consider the measurement error inherent in the test stimulus and may be misleading. Locke et al. [25] has described the calculation of a meaningful CPM effect as a percentage change from baseline (increase in pain threshold or decrease in pain ratings) greater than the inherent measurement error. In this review, judging from the reported value for CPM effect and the standard deviation, it is clear that there are differences in the response to the various CPM paradigms with some participants demonstrating inhibition of pain and others demonstrating facilitation. While some investigators have described "non-responders", this reporting is not standardized and requires improvement for transparency. While the consideration of measurement error in the calculation of a clinically meaningful effect is new to CPM studies, it is statistically robust and widely used for the interpretation of change scores [34; 36]. This approach may aide the interpretation of results across studies.

Important Findings Regarding CPM Test Design

Following exposure to a CPM conditioning stimulus, it is unclear how long pain inhibition persists. While it may be stimulus dependent, pain inhibition secondary to cold water immersion continues 10 minutes after removal of the conditioning stimulus but has resolved at 15 minutes [24]. The time for resolution of inhibition has important implications for intra-session reliability studies and studies investigating multiple pain measures.

Cold water immersion was the most frequently reported conditioning stimulus in this review, however stimulus parameters vary. Olesen et al. [32] used cold water immersion at 2° C for 3 minutes as a conditioning stimulus and reported that the majority of patients were unable to remain in the conditioning stimulus for 3 minutes due to the intensity of pain, suggesting these may be inappropriate parameters for patients with a painful condition. In this study the reliability of the CPM effect was poor (ICC= 0.10) possibly due to random error introduced by systematic differences in exposure to the conditioning stimulus.

The choice of outcome measure or response has important implications for CPM reliability. Static measures of pressure pain threshold, or the point where stimulation just becomes painful, demonstrate good to excellent reliability and in contrast, when statically measuring pressure pain tolerance, or the point when the painfulness of stimulation just becomes intolerable, re-test reliability is poor to fair [32]. Similarly, a difference is seen in the outcome or response measure to contact heat with the individualised temperature of the contact heat pain test stimulus demonstrating fair to excellent reliability, while the pain ratings for exposure to contact heat range from poor to fair.

There is evidence for gender differences in CPM effect. Martel et al. [26] investigated CPM in patients with back pain, assessing the influence of demographics including age, gender, medication use, pain severity and psychological factors including catastrophising and negative affect. They reported gender differences for the magnitude and stability of the CPM effect however with regards to demographic and psychological variables there was no significant association with CPM magnitude or stability and gender. This was supported by Valencia et al. [39] in an investigation of the influence of shoulder pain intensity and gender on CPM stability in pre- and post-surgical shoulder pain patients and in healthy volunteers with exercise induced shoulder pain. They found while the reliability of CPM was not related to shoulder pain intensity in either group, the reliability of the CPM effect differed between genders with female patients and male healthy volunteers demonstrating greater reliability.

Objective measures such as pain elicited reflexes are appealing as test stimuli for their potential to decrease subjectivity and random error and therefore to improve reliability. Biurrun Manresa et al. [3] and Jurth et al. [19] investigated the intersession reliability of CPM in healthy volunteers using the nociceptive withdrawal or flexion reflex as an objective, reliable measure of spinal nociceptive processing [4] as a test stimulus. Biurrun Manresa et al. [4] reported excellent reliability for the repeatability of the pain elicited reflex test stimulus, whereas the reliability of the cold water immersion induced CPM effect was poor. In contrast, Jurth et al. [19] reported good reliability for the hot water induced CPM effect. These results suggest the pain elicited reflex may be a reliable test stimulus, and the difference in the reliability of the CPM effect in the two studies may be secondary to the parameters of the conditioning stimulus. The pain-elicited reflex may be found to increase the objectivity and reliability of the CPM paradigm and warrants further investigation in other populations and in combination with other noxious conditioning stimuli.

As standardization in the testing of CPM is lacking, it is important to consider novel test paradigms. Granovsky et al. [14] investigated the reliability of CPM in healthy volunteers using a protocol which was novel for introducing the second test stimulus prior to rather than following the introduction of the conditioned stimulus. The intersession reliability was reported as fair (ICC = .59) however it is possible that in using a pre-determined value for tonic heat pain as a conditioning stimulus habituation to temperature may occur, with the intensity of the conditioning stimulus dropping below that necessary to induce CPM in some subjects [13]. While the single-test stimulus paradigm is enticing for the reduction in testing time, further reliability studies including an investigation of response stability are warranted.

Whilst work is required to standardise the evaluation and interpretation of CPM as an experimental and clinical measure, it is apparent that CPM has great potential as a clinically important measure or biomarker.   In a systematic review and meta-analysis, Lewis et al. [24] appraised the risk of bias and synthesised the evidence from 30 studies comparing CPM between chronic pain populations and control groups. They reported that nearly 70% of comparisons revealed a statistically significant reduction in CPM in chronic pain patients and an acceptable level of bias in included studies, providing good evidence that patients with chronic pain conditions have a significantly reduced CPM effect as compared to healthy individuals. In surgical populations, it has been reported that patients with less efficient CPM are at greater risk of developing chronic post-operative pain [40; 43] and that CPM may be predictive of subsequent pain relief (Wilder- Smith, personal communication). In pharmacological studies, it has been demonstrated that in patients with painful diabetic neuropathy, CPM predicts the analgesic effectiveness of duloxetine [44] and tapentadol (Niesters et al, personal communication) and can be activated by tapentadol [30].

While it appears that CPM is often deficient in patients with chronic pain conditions, it is unclear to what degree deficient endogenous pain modulation may be a cause or an effect of the chronic pain condition. Emerging evidence suggests that deficient CPM may be the result of a chronic pain condition, whether that pain be neuropathic or nociceptive in nature, and that when pain is alleviated, CPM is restored. This restoration or rescue of CPM has been demonstrated with the pharmacological treatment of pain [30; 44] and following joint replacement surgery in patients with painful hip osteoarthritis [20] and painful knee osteoarthritis [15].

Questions persist as well as to the nature of CPM as a stable trait or a transient state and as to how CPM is influenced by environment and context. While it is known from animal studies that DNIC in the rat can function independently of cortical control, it is unclear in humans how the descending modulation of pain may be cognitively confounded [2]. It may be that patients with chronic pain have difficulty disengaging from their pain toward a distracting stimulus, or that psychological factors such as anxiety or hyper-vigilance interfere with the pain inhibition response [2]. It has been demonstrated in humans that cognitive manipulation can effect CPM; pain inhibition under CPM appears to depend on the perceived level of the conditioned stimulus pain rather than solely on its physical intensity [31]. Additionally, in humans, there is evidence to support an association of mood and affect with CPM. In a double-blind placebo controlled randomized trial of intranasal oxytocin, Goodin et al. [12] demonstrated that oxytocin augmented CPM and reduced negative mood and anxiety.

There is evidence to suggest much potential for CPM to serve as a useful prognostic factor and predictor of response to therapeutic intervention in patients with chronic and neuropathic pain. As such, the evaluation of CPM may aid clinical decision making, assist in informing patients about possible outcomes, be used to identify risk groups for stratified management, and be a potentially modifiable target [17]. However, for a measure such as CPM to be a clinically useful prognostic factor, it must produce consistent results with minimal measurement error, i.e. it must be reliable. Estimating the reliability of CPM presents a challenge because just as there has been much heterogeneity in the investigations of CPM testing paradigms, variability in the analysis and reporting of the reliability of CPM has been equally heterogeneous.

Review Limitations

No meta-analysis was performed, therefore our findings regarding the reliability of CPM amount to a qualitative synthesis of the evidence. Additionally, our findings are limited by the quality of reporting in the included studies. While we attempted to control for the induction of reviewer bias by relying upon double screening of studies, data extraction and assessment of risk of bias, the risk of reviewer bias is nonetheless a consideration.

Conclusions

There is evidence to suggest that CPM is a reliable measure, however the degree of reliability is dependent upon stimulation parameters, study methodology and the population of interest. The validation of CPM as a robust prognostic factor in experimental and clinical pain studies will be facilitated by improvements in the reporting of CPM reliability studies.

Recommendations for Future Research

It has been recommended that the CPM effect should be reported as both the absolute change and the percent change (when appropriate for the level of measurement) in the perceived test stimulus induced by the conditioning stimulus and a measure of variability should be included [41]. Recommendations for future reliability studies include due consideration of how the results for a sample of participants may be generalized to a population of interest. Gierthmuhlen et al. [11] has described important data collection domains for healthy volunteer quantitative sensory testing studies which may be equally pertinent for dynamic measures such as CPM, including but not limited to socio-demographic data, medical history and current health status, pain coping strategies, psychological factors, history of alcohol and drug abuse, smoking and use of recreational drugs, current medication, depression and anxiety scores, the frequency of any pain episodes during the last 3-6 months and self-reported sleep measurements. Consideration should be given to blinding of both the investigator and the participants of CPM studies, standardization of test instructions and as to how the test environment and exposure to investigators and other study participants may bias performance or results. The intensity and exposure time for the conditioning stimulus should be of a magnitude that the stimulus is uniform for all participants. Attempts to control for known confounders should be made, with an accounting of confounders at both baseline and retest. Lastly, improvements in the statistical design and analysis of CPM reliability studies are essential if progress is to be made toward standardization in CPM testing and reporting. The inclusion of a sample size calculation, an appropriate reliability coefficient and 95% confidence interval and a measure of response stability will aid the interpretation of results and the comparison between studies. Thorough data reporting including measures of central tendency and variability for ratings for test stimulus, conditioning stimulus, conditioned test stimulus and CPM magnitude, the number of responders and non-responders and how this was established, the intra or intersession reliability for the test and conditioning stimulus, and where appropriate the absolute and percentage change for the CPM effect will aide comparison of testing paradigms across studies and substantiate the repeatability and inherent variability of the CPM paradigm.

Acknowledgements

This report is independent research and the views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. Donna Kennedy is funded by a National Institute for Health Research and

Health Education England Clinical Doctoral Research Fellowship. Dr Harriet Kemp is funded by a European Commission NeuroPain FP7 Grant EC (#2013-602891). Ms Deborah Ridout has received financial support from the National Institute for Health Research and Health Education England for her contribution to this work. Professor David Yarnitsky and Professor Andrew SC Rice have received financial support from DOLORisk, a European Union Horizon 2020 research and innovation programme (grant 633491) for their contribution to this manuscript.

Conflicts of Interest

The authors declare they have no conflicts of interest to report.

References

[1] Baumgartner TA. Norm-referenced measurement: Reliability. Champaign, IL, 1989.

[2] Bingel U, Tracey I. Imaging CNS modulation of pain in humans. Physiology 2008;23:371-380; http://www.ncbi.nlm.nih.gov/pubmed/19074744.

[3] Biurrun Manresa JA, Fritsche R, Vuilleumier PH, Oehler C, Morch CD, Arendt-Nielsen L, Andersen OK, Curatolo M. Is the conditioned pain modulation paradigm reliable? A test-retest assessment using the nociceptive withdrawal reflex. PloS one 2014;9(6):e100241; http://www.ncbi.nlm.nih.gov/pubmed/24950186.

[4] Biurrun Manresa JA, Neziri AY, Curatolo M, Arendt-Nielsen L, Andersen OK. Test-retest reliability of the nociceptive withdrawal reflex and electrical pain thresholds after single and repeated stimulation in patients with chronic low back pain. Eur J Appl Physiol 2011;111(1):83-92; http://www.ncbi.nlm.nih.gov/pubmed/20814801.

[5] Bruton A, Conway JH, Holgate ST. Reliability: What is it and how is it measured? Physiotherapy 2000;86(2):94-99;

[6] Cathcart S, Winefield AH, Rolan P, Lushington K. Reliability of temporal summation and diffuse noxious inhibitory control. Pain Res Manag 2009;14(6):433-438; http://www.ncbi.nlm.nih.gov/pubmed/20011713.

[7] Coghill RC, Yarnitsky D. Healthy and normal? The need for clear reporting and flexible criteria for defining control participants in quantitative sensory testing studies. Pain 2015; http://www.ncbi.nlm.nih.gov/pubmed/26313407.

[8] Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature 2014;505(7485):612-613; http://www.ncbi.nlm.nih.gov/pubmed/24482835.

[9] Edwards RR, Ness TJ, Weigent DA, Fillingim RB. Individual differences in diffuse noxious inhibitory controls (DNIC): association with clinical variables. Pain 2003;106(3):427-437;

[10] Ge HY, Madeleine P, Arendt-Nielsen L. Sex differences in temporal characteristics of descending inhibitory control: an evaluation using repeated bilateral experimental induction of muscle pain. Pain 2004;110(1-2):72-78; http://www.ncbi.nlm.nih.gov/pubmed/15275754.

[11] Gierthmuhlen J, Enax-Krumova EK, Attal N, Bouhassira D, Cruccu G, Finnerup NB, Haanpaa M, Hansson P, Jensen TS, Freynhagen R, Kennedy JD, Mainka T, Rice A, Segerdahl M, Sindrup SH, Serra J, Tolle T, Treede RD, Baron R, Maier C. Who is healthy? Aspects to consider when including healthy volunteers in QST-based studies- a consensus statement by the EUROPAIN and NEUROPAIN consortia. Pain 2015; http://www.ncbi.nlm.nih.gov/pubmed/26075963.

[12] Goodin BR, Anderson AJB, Freeman EL, Bulls HW, Robbins MT, Ness TJ. Intranasal Oxytocin Administration is Associated With Enhanced Endogenous Pain Inhibition and Reduced Negative Mood States. The Clinical journal of pain 2015;31(September):757-767;

[13] Granot M, Weissman-Fogel I, Crispel Y, Pud D, Granovsky Y, Sprecher E, Yarnitsky D. Determinants of endogenous analgesia magnitude in a diffuse noxious inhibitory control (DNIC) paradigm: do conditioning stimulus painfulness, gender and personality variables matter? Pain 2008;136(1-2):142-149; http://www.ncbi.nlm.nih.gov/pubmed/17720319.

[14] Granovsky Y, Miller-Barmak A, Goldstein O, Sprecher E, Yarnitsky D. CPM Test-Retest Reliability: "Standard" vs "Single Test-Stimulus" Protocols. Pain Med 2015; http://www.ncbi.nlm.nih.gov/pubmed/26272736.

[15] Graven-Nielsen T, Wodehouse T, Langford RM, Arendt-Nielsen L, Kidd BL. Normalization of widespread hyperesthesia and facilitated spatial summation of deep-tissue pain in knee osteoarthritis patients after knee replacement. Arthritis and rheumatism 2012;64(9):2907-2916; http://www.ncbi.nlm.nih.gov/pubmed/22421811.

[16] Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med 2006;144(6):427-437; http://www.ncbi.nlm.nih.gov/pubmed/16549855.

[17] Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med 2013;158(4):280-286; http://www.ncbi.nlm.nih.gov/pubmed/23420236.

[18] Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration, 2011.

[19] Jurth C, Rehberg B, von Dincklage F. Reliability of subjective pain ratings and nociceptive flexion reflex responses as measures of conditioned pain modulation. Pain Res Manag 2014;19(2):93-96; http://www.ncbi.nlm.nih.gov/pubmed/24555177.

[20] Kosek E, Ordeberg G. Abnormalities of somatosensory perception in patients with painful osteoarthritis normalize following successful treatment. Eur J Pain 2000;4(3):229-238; http://www.ncbi.nlm.nih.gov/pubmed/10985866.

[21] Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, Finkelstein R, Fisher M, Gendelman HE, Golub RM, Goudreau JL, Gross RA, Gubitz AK, Hesterlee SE, Howells DW, Huguenard J, Kelner K, Koroshetz W, Krainc D, Lazic SE, Levine MS, Macleod MR, McCall JM, Moxley RT, 3rd, Narasimhan K, Noble LJ, Perrin S, Porter JD, Steward O, Unger E, Utz U, Silberberg SD. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 2012;490(7419):187-191; http://www.ncbi.nlm.nih.gov/pubmed/23060188.

[22] Le Bars D, Dickenson AH, Besson JM. Diffuse noxious inhibitory controls (DNIC). I. Effects on dorsal horn convergent neurones in the rat. Pain 1979;6(3):283-304; http://www.ncbi.nlm.nih.gov/pubmed/460935.

[23] Lewis GN, Heales L, Rice DA, Rome K, McNair PJ. Reliability of the conditioned pain modulation paradigm to assess endogenous inhibitory pain pathways. Pain Res Manag 2012;17(2):98-102;

[24] Lewis GN, Rice DA, McNair PJ. Conditioned pain modulation in populations with chronic pain: a systematic review and meta-analysis. Journal of Pain 2012;13(10):936-944;

[25] Locke D, Gibson W, Moss P, Munyard K, Mamotte C, Wright A. Analysis of meaningful conditioned pain modulation effect in a pain-free adult population. The journal of pain : official journal of the American Pain Society 2014;15(11):1190-1198; http://www.ncbi.nlm.nih.gov/pubmed/25241218.

[26] Martel MO, Wasan AD, Edwards RR. Sex differences in the stability of conditioned pain modulation (CPM) among patients with chronic pain. Pain Med 2013;14(11):1757-1768; http://www.ncbi.nlm.nih.gov/pubmed/23924369.

[27] Millan MJ. Descending control of pain. Progress in neurobiology 2002;66(6):355-474; http://www.ncbi.nlm.nih.gov/pubmed/12034378.

[28] Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Journal of clinical epidemiology 2009;62(10):1006-1012; http://www.ncbi.nlm.nih.gov/pubmed/19631508.

[29] Nahman-Averbuch H, Yarnitsky D, Granovsky Y, Gerber E, Dagul P, Granot M. The role of stimulation parameters on the conditioned pain modulation response. Scandanavian Journal of Pain 2013;4(1):10-14;

[30] Niesters M, Proto PL, Aarts L, Sarton EY, Drewes AM, Dahan A. Tapentadol potentiates descending pain inhibition in chronic pain patients with diabetic polyneuropathy. British journal of anaesthesia 2014;113(1):148-156; http://www.ncbi.nlm.nih.gov/pubmed/24713310.

[31] Nir RR, Yarnitsky D, Honigman L, Granot M. Cognitive manipulation targeted at decreasing the conditioning pain perception reduces the efficacy of conditioned pain modulation. Pain 2012;153(1):170-176; http://www.ncbi.nlm.nih.gov/pubmed/22119318.

[32] Olesen SS, van Goor H, Bouwense SA, Wilder-Smith OH, Drewes AM. Reliability of static and dynamic quantitative sensory testing in patients with painful chronic pancreatitis. Regional anesthesia and pain medicine 2012;37(5):530-536; http://www.ncbi.nlm.nih.gov/pubmed/22854397.

[33] Oono Y, Hongling N, Lima Matos R, Wanga K, Arendt-Nielsen L. The inter- and intra-individual variance in descending pain modulation evoked by different conditioning stimuli in healthy men. Scandinavian journal of pain 2011;2(4):162–169;

[34] Portney LG, Watkins MP. Foundations of Clinical Research. New Jersey: Prentice Hall Health, 2000.

[35] Pud D, Granovsky Y, Yarnitsky D. The methodology of experimentally induced diffuse noxious inhibitory control (DNIC)-like effect in humans. Pain 2009;144(1-2):16-19; http://www.ncbi.nlm.nih.gov/pubmed/19359095.

[36] Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clinical rehabilitation 1998;12(3):187-199; http://www.ncbi.nlm.nih.gov/pubmed/9688034.

[37] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420-428; http://www.ncbi.nlm.nih.gov/pubmed/18839484.

[38] Valencia C, Fillingim RB, Bishop M, Wu SS, Wright TW, Moser M, Farmer K, George SZ. Investigation of Central Pain Processing in Post-Operative Shoulder Pain and Disability. The Clinical journal of pain 2013; http://www.ncbi.nlm.nih.gov/pubmed/24042347.

[39] Valencia C, Kindler LL, Fillingim RB, George SZ. Stability of conditioned pain modulation in two musculoskeletal pain models: investigating the influence of shoulder pain intensity and gender. BMC musculoskeletal disorders 2013;14(1):182; http://www.ncbi.nlm.nih.gov/pubmed/23758907.

[40] Wilder-Smith OH, Schreyer T, Scheffer GJ, Arendt-Nielsen L. Patients with chronic pain after abdominal surgery show less preoperative endogenous pain inhibition and more postoperative hyperalgesia: a pilot study. J Pain Palliat Care Pharmacother 2010;24(2):119-128; http://www.ncbi.nlm.nih.gov/pubmed/20504133.

[41] Yarnitsky D, Arendt-Nielsen L, Bouhassira D, Edwards RR, Fillingim RB, Granot M, Hansson P, Lautenbacher S, Marchand S, Wilder-Smith O. Recommendations on terminology and practice of psychophysical DNIC testing. Eur J Pain 2010;14(4):339; http://www.ncbi.nlm.nih.gov/pubmed/20227310.

[42] Yarnitsky D, Bouhassira D, Drewes AM, Fillingim RB, Granot M, Hansson P, Landau R, Marchand S, Matre D, Nilsen KB, Stubhaug A, Treede RD, Wilder-Smith OH. Recommendations on practice of conditioned pain modulation (CPM) testing. Eur J Pain 2015;19(6):805-806; http://www.ncbi.nlm.nih.gov/pubmed/25330039.

[43] Yarnitsky D, Crispel Y, Eisenberg E, Granovsky Y, Ben-Nun A, Sprecher E, Best LA, Granot M. Prediction of chronic post-operative pain: pre-operative DNIC testing identifies patients at risk. Pain 2008;138(1):22-28; http://www.ncbi.nlm.nih.gov/pubmed/18079062.

[44] Yarnitsky D, Granot M, Nahman-Averbuch H, Khamaisi M, Granovsky Y. Conditioned pain modulation predicts duloxetine efficacy in painful diabetic neuropathy. Pain 2012;153(6):1193-1198; http://www.ncbi.nlm.nih.gov/pubmed/22480803.

Appendix A. Search strategy

1. AMED, EMBASE, Medline, CINAHL; (Conditioned AND pain AND modulation).ti,ab; 565 results.
2. AMED, EMBASE, Medline, CINAHL; (diffuse AND noxious AND inhibitory AND control).ti,ab; 345 results.
3. AMED, EMBASE, Medline, CINAHL; DNIC.ti,ab; 895 results.
4. AMED, EMBASE, Medline, CINAHL; (Heterotopic AND noxious AND conditioning).ti,ab; 147 results.
6. AMED, EMBASE, Medline, CINAHL; 1 OR 2 OR 3 OR 4; 1589 results.
7. AMED, EMBASE, Medline, CINAHL; reliability.ti,ab; 270751 results.
8. AMED, EMBASE, Medline, CINAHL; repeatability.ti,ab; 36159 results.
9. AMED, EMBASE, Medline, CINAHL; stability.ti,ab; 596078 results.
10. AMED, EMBASE, Medline, CINAHL; 7 OR 8 OR 9; 887198 results.
11. AMED, EMBASE, Medline, CINAHL; 6 AND 10; 69 results.
12. AMED,EMBASE,Medline,CINAHL; Duplicate filtered: [6 AND 10]; 69 results.

Appendix B. Full-text study exclusion

Reference	Reason for exclusion
O'Neill et al. [30] "Reliability and validity of a simple and clinically applicable pain stimulus: Sustained mechanical pressure with a spring-clamp". *Chiropractic and Manual Therapies*, 22/1.	Not a CPM study

Appendix C.

The QUIPS Tool domains (Hayden et al 2006) operationalized **(bold)** for the evaluation of the repeatability of a CPM test paradigm.

1. Study participation considers the proportion of eligible persons who participate in the study, descriptions of the source population, baseline study sample, sampling frame and recruitment, and adequate inclusion and exclusion criteria including explicit diagnostic criteria.
2. Study attrition addresses whether participants with follow-up data **(re-test data)** represent persons enrolled in the study or was the outcome biased by a selective group who completed the study.
3. Prognostic factor measurement domain assists in determining if the prognostic factor was measured in a similar and valid way for all participants. This includes items pertinent to internal validity including investigator and participant blinding and measurement methods. **Risk of bias was rated as low where the conditioning stimulus was consistent between participants and where information is provided regarding participant blinding (i.e. blinding to the intention of study; use of a script for consistency in test instructions between participants; information regarding participant exposure during test interval). Risk of bias was moderate where one factor was reported, high were neither factor was reported.**
4. Outcome measurement considers whether outcome was measured in a valid and reliable way for all participants, for example, with **a validated pain scale or measure**.
5. Study confounding aids the assessor in judging whether another confounding factor may explain the reported association between the factor of interest and outcome. To make this

judgment, the assessor considers the measurement of potential confounders and whether all important confounding factors are accounted for in the study design or analysis. **For risk of bias in confounding in this review, risk was rated as low where at least 4 confounders were accounted for at baseline *and* re-test; moderate risk where 3 are accounted for and high risk for less than 3 [22]. In healthy volunteers, potential confounders may include but are not limited to the presence or level of pain prior to testing, screening for conditions which may affect pain threshold (i.e. chronic pain conditions such as fibromyalgia; peripheral neuropathy), oestrus cycle, caffeine and medication intake prior to testing, psychological factors including anxiety and depression [11], time of day and exercise. Additional patient group confounders may include medical diagnosis based on accepted criteria, stable regime of pain treatment and pharmacologic treatment and no additional painful condition other than the investigated diagnosis [22]. In addition to accounting for confounding factors at the initiation of the study or at baseline measures, it is essential that the potentially confounding factor is accounted for at the time of re-test.**

6. Statistical analysis and reporting addresses the appropriateness of the study's statistical analysis and thoroughness of reporting [17] with the aim of insuring that an appropriate design and adequate reporting limit the possibility for the presentation of invalid or spurious results. **Three important elements of statistical design for reliability studies include a sample size calculation, appropriate reliability coefficient and 95% confidence interval and the reporting of sufficient data to allow for the assessment of the adequacy of the analysis with no selective reporting of results. Risk of bias was rated as low were all three components were reported, moderate where two components were reported and high where one component was reported.**

Table 1. Data summary and results of included CPM reliability studies. Healthy volunteer (HV); Cold pressor test (CPT); ischemic pain (IP); pressure pain threshold (PPT); pressure pain tolerance (PPTol); hot water bath (HWB); nociceptive withdrawal reflex (NWR); nociceptive flexion reflex (NFR); intraclass correlation coefficient (ICC); coefficient of repeatability (CR); intra-individual stability coefficient (ISC); coefficient of variation (CV); minimal detectable change (MDC); standard error of measurement (SEM); standard error (SE). ***Bold, italicized data*** is not reported***.***

Table 2. Protocol violations in the administration of the test and conditioning stimuli and reliability of test and conditioned stimuli across test sessions. Cold pressor test (CPT); ischemic pain (IP); pressure pain threshold (PPT); pressure pain tolerance (PPTol); hot water bath (HWB); nociceptive withdrawal reflex (NWR); nociceptive flexion reflex (NFR)); intraclass correlation coefficient (ICC). ***Bold, italicized data*** is not reported***.***

Table 3. Risk of bias in CPM reliability studies (Hayden et al 2006, Hayden et al 2013).

Figure 1. Study flow diagram.

Figure 2. Risk of bias in study confounding and reliability. ICC is the highest reported reliability coefficient for CPM effect. For risk of bias score, 1 = low risk; 2 = moderate risk; 3 = high risk.

Table 1. Demographics, CPM paradigm and reliability results. Male/female (M/F); standard deviation (SD); healthy volunteer (HV); confidence interval (CI); cold pressor test (CPT); ischemic pain (IP); pressure pain threshold (PPT); pressure pain tolerance (PPTol); hot water bath (HWB); nociceptive withdrawal reflex (NWR); nociceptive flexion reflex (NFR); intraclass correlation coefficient (ICC); coefficient of repeatability (CR); intra-individual stability coefficient (ISC); coefficient of variation (CV); minimal detectable change (MDC); standard error of measurement (SEM); standard error (SE). ***Bold, italicized*** not reported***.***

Study	**Sample size (M/F)**	**Population Age mean (SD)**	**Test stimulus Test site**	**Conditioning stimulus (paradigm)**	**Re-test interval**	**Reliability coefficient (95% CI)**	**Response Stability**
Cathcart et al. [7]	20 (9/11)	HV; male 27 (6.4), female 23 (3.6)	PPT 1. Right middle finger 2. Right trapezius	IP left arm (parallel)	Intra-session	1. finger ICC = 0.57 2. shoulder ICC = 0.69 ***95% CI not reported***	CR= 0.35 (±1.69)
Oono et al. [32]	12 (12/0)	HV; 25.6 ± 1.5 (SEM)	PPT, PPTol 1.Right masseter muscle 2. Left forearm 3.Left tibialis anterior	1.CPT hand 2.IP upper arm 3. Pressure pain- head band (parallel)	2 days	***Not reported***	Inter-individual CV =41.4%; intra-individual CV=40.1% for CPM effect with CPT as CS, PPTol at forearm as TS
Lewis et al. [21]	22 (7/15)	HV; 25(8)	PPT Medial right knee	1.CPT left hand 2.IP left arm (parallel)	Intra- session; 3 days	Intra-session: CPT ICC= 0.85 (0.62–0.94); IP ICC=0.75 (0.35–0.90). Intersession: CPT ICC= 0.66 (0.12–0.87); IP ICC=−0.4 (−1.8–0.4)	***Not reported***
Olesen et al. [31]	62 (38/24)	Painful chronic pancreatitis; 53 (11)	PPT Quadriceps	CPT right hand (sequential)	1 week	ICC= 0.10 ***95% CI not reported***	***Not reported***
Martel et al. [24]	55 (35/20)	Chronic back pain; Men 48.9 (10.5) Women 49.5 (8.9)	PPT Right trapezius	CPT left hand (parallel)	10 days	Overall sample: ICC = 0.59 (0.38-0.74); Women: ICC = 0.75 (0.56-0.87); Men: ICC = 0.33 (0.12-0.67)	Men ISC= 0.29; women ISC= 0.79; overall ISC= 0.61
Valencia et al. [38]	HV 190 (74/116) Patients 134 (87/47)	HV, shoulder pain patients; HV 23.02 (6.04), patients 43.83 (17.8)	Contact heat pain (50/100) Thenar eminence	CPT contralateral hand (sequential)	HV-intra-session & 1,3,5 days Patients intra-session; pre-surgery, 3 months post- surgery	Intra-session, patients, pre-surgery ICC= 0.54 (0.34-0.68); post-op ICC= 0.62 (0.43-0.74). Intra-session HV; ICC =0.66 (0.55-0.75)- ICC = 0.72 (0.62-0.79).Inter-session, HVs; female ICC = 0.65 (0.51-0.75); male ICC= 0.82 (0.73-0.88)	Female patients, intra-session pre-surgery SEM=5.83; post- surgery SEM= 4.25. Male patients, intra-session pre-surgery SEM= 7.33; post-surgery SEM= 6.50
Wilson et al. [40]	22 (0/22)	HV; 27 (7)	Contact heat pain (6/10) Dominant forearm	HWB (46.5° C) non-dominant hand (parallel)	Repeated 8 times over 4 menstrual cycles	ICC= 0.39 (0.23-0.59)	Estimated marginal grand mean ± SE = 1.3 ± 0.3

Biurrun Manresa et al. [4]	34 (34/0)	HV; 27.5 (6.8)	1.NWR threshold at biceps and rectus femoris
			2.Electrical pain detection threshold
			3. pain intensity rating electrical stimulation
				CPT
				contralateral hand (parallel)
					Between 1-3 weeks (average 11.9 ± 1.9 days)	1. NWR threshold ICC= 0.26 (0-0.55);
						2. Electrical pain detection threshold ICC= 0.09 (0-0.41)
						3. Pain intensity ratings ICC = 0.44 (0.13-0.68).
							NWR threshold:
							Bland-Altman analysis bias = 0.3;
							LoA = -5.4–6.0;
							CV (95% CI) = 64.1% (39.1%–81.8%).

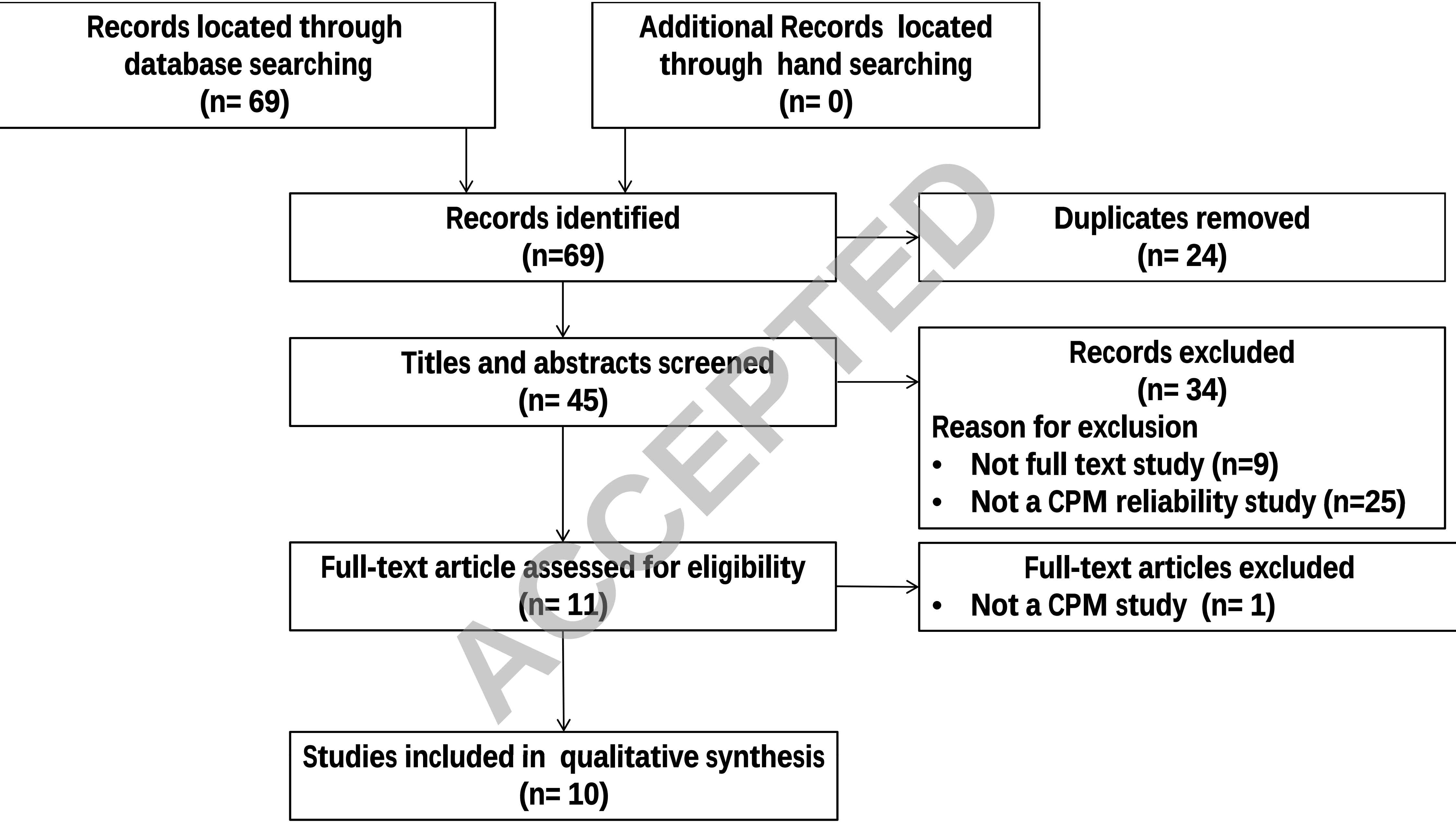
Jurth et al. [18]	40 (20/20)	HV; ***Not reported***	1. NFR biceps femoris (pain 50/100) 2.Subjective pain ratings (0-100 NRS)	HWB (parallel)	28 days	CPM effect with NFR ICC= 0.61 (0.36-0.78). Subjective pain ratings for CPM effect ICC= 0.54 (0.26-0.74).	***Not reported***
Granovsky et al. [14]	1) 35 (10/25) 2a+b) 30 (15/15)	HV; 1) 26.1 (2.5) 2a+b) 25.9 (2.6)	1)Contact heat pain, 60/100 dominant hand 2a) 2 thermode + 2b) single test stimulus- contact heat pain, 30/100 non-dominant volar forearm	1) HWB dominant hand 2a+b) contact heat pain dominant upper arm (parallel)	1) 3-7 days 2a+b) 7 days	CPM effect 1)ICC = 0.34 (0.03–0.59) 2a) ICC= 0.21 (-0.15 to 0.53) 2b) ICC= 0.59 (0.30–0.78)	***Not reported***

Table 2. Protocol violations and stimulus reliability. Cold pressor test (CPT); ischemic pain (IP); pressure pain threshold (PPT); pressure pain tolerance (PPTol); hot water bath (HWB); nociceptive withdrawal reflex (NWR); nociceptive flexion reflex (NFR)); intraclass correlation coefficient (ICC); not reported (***NR)***

	TS protocol violations	**CS protocol violations**	**TS Test-retest reliability**	**CS Test-retest reliability**
Cathcart et al.[7]	PPT- ***NR***	IP- ***NR***	PPT Intra-session ICC= 0.82	***NR***
Oono et al. [32]	PPT, PPTol - ***NR***	CPT 2-4° C, 10 minutes- most participants did not tolerate on first attempt. IP, mechanical pressure- ***NR***	***NR***	***NR***
Lewis et al. [21]	PPT- ***NR***	CPT 12 ± 1° C, 2 minutes, IP- ***NR***	PPT intra-session ICC= 0.87 (0.60-0.95) PPT intersession ICC= 0.65 (0.05-0.87)	IP NPS intra-session ICC= 0.60 (0.24–0.82), intersession ICC=0.82 (0.59–0.92). CPT NPS intra-session ICC= 0.94 (0.86–0.98), intersession ICC= 0.80 (0.56–0.92)
Olesen et al. [31]	PPT quadriceps- ***NR***	CPT 2° C, 3 minutes- tolerated for median of 38 seconds at baseline, 35 seconds on retest	PPT intersession ICC = 0.79	***NR***
Martel et al. [24]	PPT trapezius- ***NR***	CPT 4°C, 2 minutes- ***NR***	PPT intersession ICC= 0.72 (0.56–0.83)	CPT pain ratings ICC = 0.61 (0.41–0.75)
Valencia et al. [38]	Contact heat pain (50/100) - ***NR***	CPT 8° C, 1 minute- ***NR***	***NR***	***NR***
Wilson et al. [40]	Contact heat pain (6/10) ***NR***	HWB 46.5° C, 1 minute- ***NR***	Temperature °C intersession ICC=0.83 (0.72–0.91) VNPS intersession ICC= 0.40 (0.24–0.60)	HWB VNPS ICC = 0.79 (0.68–0.89)
Biurrun Manresa et al. [4]	1. NWR threshold- ***NR***; unable to elicit NWR in 5 subjects (13%) 2.Electrical pain detection threshold- ***NR*** 3. pain rating- electrical stimulation- ***NR***	CPT <2°C, 2 minutes or until reaching 7/10 on VAS - 4 of 34 (12%) of subjects did not tolerate continuously	Intersession 1.NWR ICC= 0.93 (0.87–0.97) 2. Electrical pain detection threshold ICC=0.67 (0.43–0.82) 3. Pain intensity ratings ICC= 0.85 (0.71–0.92)	***NR***
Jurth et al. [18]	NFR- ***NR***	HWB 46.5°C, 200 seconds-***NR***, 1 subject excluded	***NR***	***NR***
Granovsky et al. [14]	1.Contact heat Pain ^{60}- ***NR*** 2a,b. Contact heat, Pain^{30}- ***NR***	1.HWB 46.5° C, 1 minute- ***NR*** 2a,b.Contact heat, TS + .5° C- ***NR***	Intersession 1. Bath- thermode contact heat °C ICC= 0.53; contact heat NPS ICC= 0.31 2a. 2 thermode contact heat pain °C ICC=0.64; mean NPS ICC= 0.19 2b. Single test stimulus contact heat pain test stimulus NPS ICC = 0.15	1. Bath –thermode HWB NPS ICC= 0.76 2a.2 thermode contact heat pain VAS ICC= 0.16 2b. Single test stimulus not reported

Table 3. Risk of bias in CPM reliability studies (Hayden et al 2006, Hayden et al 2013).

Study	**Study Participation**	**Study Attrition**	**Prognostic Factor measurement**	**Outcome Measurement**	**Confounding**	**Statistical Analysis & Reporting**
Cathcart et al.(2009)	moderate	moderate	moderate	low	low	high
Oono et al. (2011)	high	moderate	moderate	low	moderate	high
Lewis et al. [21]	high	moderate	moderate	low	low	moderate
Olesen et al (2012)	moderate	moderate	high	low	high	high
Martel et al. (2013)	moderate	moderate	moderate	low	moderate	moderate
Valencia et al. (2013)	moderate	moderate	high	low	moderate	moderate
Wilson et al. (2013)	low	low	moderate	low	high	moderate
Biurrun Manresa et al. (2014)	moderate	moderate	moderate	low	high	low
Jurth et al. (2014)	high	moderate	moderate	low	low	high
Granovsky et al. (2015)	high	moderate	moderate	low	low	moderate



Records located through database searching (n= 69)

Additional Records located through hand searching (n= 0)

Records identified (n=69)

Duplicates removed (n= 24)

Titles and abstracts screened (n= 45)

Records excluded (n= 34)

Reason for exclusion

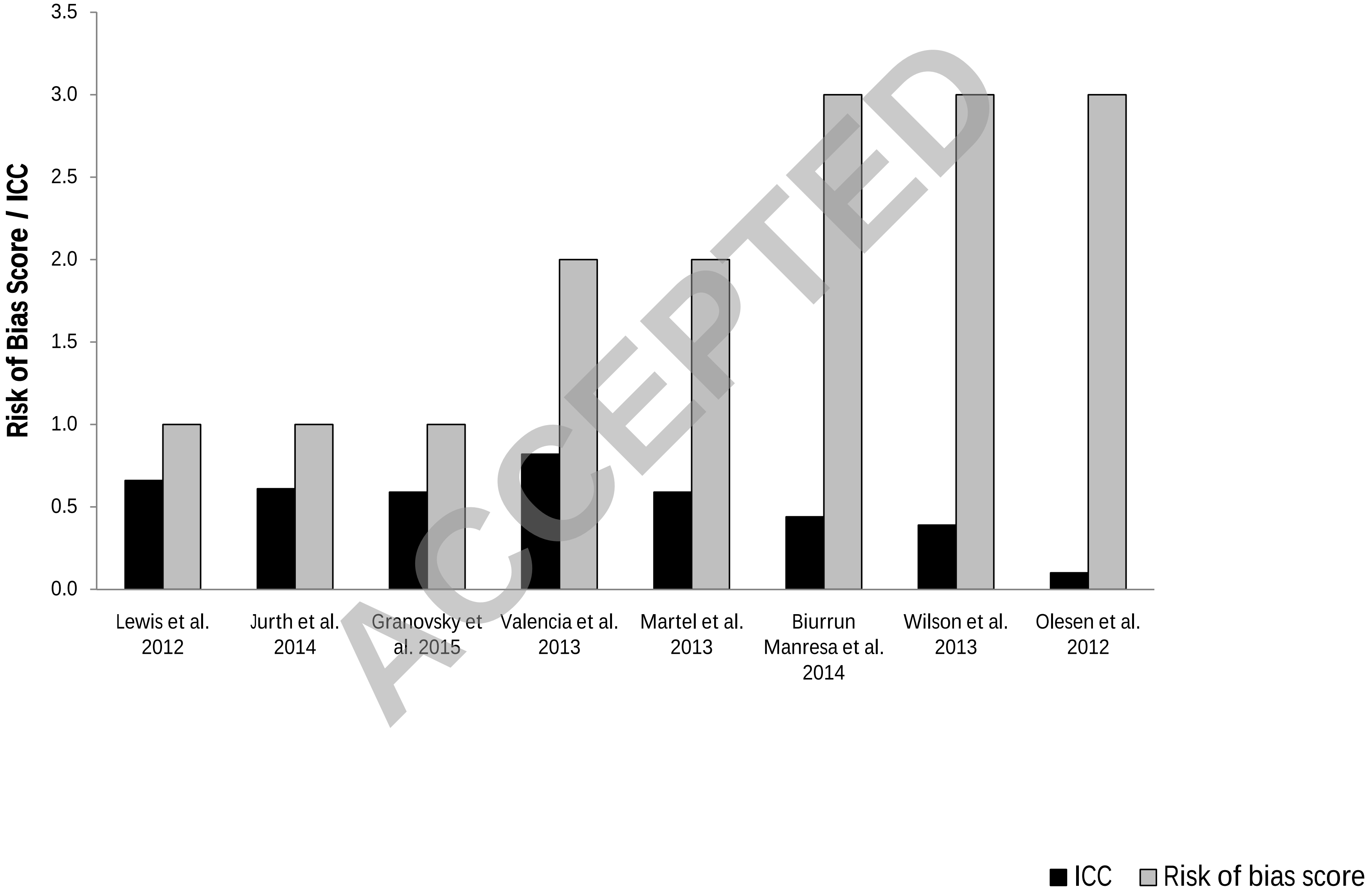
- Not full text study (n=9)
- Not a CPM reliability study (n=25)

Full-text article assessed for eligibility (n= 11)

Full-text articles excluded

- Not a CPM study (n= 1)

Studies included in qualitative synthesis (n= 10)


3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Risk of Bias Score / ICC
Lewis et al. 2012
Jurth et al. 2014
Granovsky et al. 2015
Valencia et al. 2013
Martel et al. 2013
Biurrun Manresa et al. 2014
Wilson et al. 2013
Olesen et al. 2012
■ ICC
▨ Risk of bias score