Reliability of conditioned pain modulation in healthy individuals and chronic pain patients

a systematic review and meta-analysis

Nuwailati, Rania; Bobos, Pavlos; Drangsholt, Mark; Curatolo, Michele

Published in:
Scandinavian Journal of Pain

DOI (link to publication from Publisher):
10.1515/sjpain-2021-0149

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Systematic Review

Rania Nuwailati*, Pavlos Bobos, Mark Drangsholt and Michele Curatolo

Reliability of conditioned pain modulation in healthy individuals and chronic pain patients: a systematic review and meta-analysis

https://doi.org/10.1515/sjpain-2021-0149
Received August 26, 2021; accepted January 21, 2022;
published online February 10, 2022

Abstract

Objectives: Conditioned pain modulation (CPM) is a psychophysical parameter that is used to reflect the efficacy of endogenous pain inhibition. CPM reliability is important for research and potential clinical applications. The aim of this systematic review and meta-analysis was to evaluate the reliability of CPM tests in healthy individuals and chronic pain patients.

Methods: We searched three databases for peer-reviewed studies published from inception to October 2020: EMBASE, Web of Science and NCBI. Risk of bias and the quality of the included studies were assessed. A meta-analysis with a random effects model was conducted to estimate intraclass correlation coefficients (ICCs).

Results: Meta-analysis was performed on 25 papers that examined healthy participants (k=21) or chronic pain patients (k=4). The highest CPM intra-session reliability was with pressure as test stimulus (TS) and ischemic pressure (IP) or cold pressor test (CPT) as conditioning stimulus (CS) in healthy individuals (ICC 0.64, 95% CI 0.45–0.77), and pressure as TS with CPT as CS in patients (ICC 0.77, 95% CI 0.70–0.82). The highest inter-session ICC was with IP as TS and IP or CPT as CS (ICC 0.51, 95% CI 0.42–0.59) in healthy subjects. The only data available in patients for inter-session reliability were with pressure as TS and CPT as CS (ICC 0.44, 95% CI 0.11–0.69). Quality ranged from very good to excellent using the QACMRR checklist. The majority of the studies (24 out of 25) scored inadequate in Kappa coefficient reporting item of the COSMIN-ROB checklist.

Conclusions: Pressure and CPT were the TS and CS most consistently associated with good to excellent intra-session reliability in healthy volunteers and chronic pain patients. The inter-session reliability was fair or less for all modalities, both in healthy volunteers and chronic pain patients.

Keywords: conditioned pain modulation; meta-analysis; pain assessment; pain measurement; reliability.

Introduction

Endogenous pain modulation is a broad term that describes multiple central nervous system mechanisms resulting in reducing or increasing pain [1]. The psychophysical parameter of conditioned pain modulation (CPM) is commonly used to assess the efficacy of endogenous pain inhibition in humans [2–4]. The CPM test paradigm involves the "pain inhibits pain" phenomenon, where pain with a test stimulus (TS) is reduced either during the application or immediately after discontinuation of a conditioning stimulus (CS) [2, 4]. The extent of the resulting inhibition in pain is the CPM magnitude, typically presented as positive CPM values. However, the categorization into positive or negative values may be misleading, as a recent study showed that healthy subjects can display negative CPM values [5]. It has been proposed that CPM effects should be reported on an individual level as facilitatiry and inhibitory manifestations [6].

***Corresponding author: Rania Nuwailati**, Department of Oral Health Sciences, University of Washington, Seattle, WA, USA; and Department of Oral Medicine, University of Washington, Seattle, WA, USA, E-mail: ranian@uw.edu
Pavlos Bobos, Applied Health Research Centre (AHRC), The HUB, Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, ON, Canada; and Department of Clinical Epidemiology and Health Care Research, Dalla Lana School of Public Health, Institute of Healthy Policy Management and Evaluation, University of Toronto, Toronto, ON, Canada, E-mail: p.bobos@mail.utoronto.ca
Mark Drangsholt, Department of Oral Health Sciences, University of Washington, Seattle, WA, USA; and Department of Oral Medicine, University of Washington, Seattle, WA, USA, E-mail: drangs@uw.edu
Michele Curatolo, Department of Anesthesiology and Pain Medicine, School of Medicine, University of Washington, Seattle, WA, USA; Harborview Injury Prevention and Research Center, Seattle, WA, USA; CLEAR Research Center for Musculoskeletal Disorders, Seattle, WA, USA; and Department of Health Science and Technology, Center for Sensory-Motor Interaction, Aalborg University, Aalborg, Denmark, E-mail: curatolo@uw.edu. https://orcid.org/0000-0003-4130-6826

The CPM paradigm has been used to assess pain inhibition in several clinical studies [7–10]. There is some evidence for the role of inefficient pain inhibition in promoting chronic pain development [2, 11]. Patients with different pain conditions have displayed a less efficient CPM effect when compared to healthy individuals [1, 2, 12–14]. However, other reviews reported mixed results on the CPM effect in pain and control groups [15, 16]. The CPM test was proposed to be used as a biomarker to identify individuals at high risk to develop chronic pain [11, 17]; the probability to develop persistent pain after surgery in patients with poor preoperative endogenous analgesic response was higher than patients with a sufficient endogenous analgesia [18]. In addition, CPM test was suggested to predict response to treatments that modulate endogenous inhibition [8, 14], which may improve personalized preventive treatment [1, 14, 17, 19, 20]. Kisler et al. used psychophysical pain measures that included CPM to predict duloxetine efficacy, suggesting that patients expressing a higher pain modulation profile will benefit more from duloxetine than a placebo [21]. Fernandes et al. reported nonsignificant correlations between CPM efficiency and most clinical manifestations of pain, but the studies were characterized by highly heterogeneous methodologies [22].

One essential requirement for the use of CPM is its reliability. Results of the several previous studies have ranged from poor to excellent reliability [23–27]. In an attempt to improve the CPM test reliability, studies have investigated different modalities, such as different combinations of TS and CS, anatomic sites, and stimulus intensities.

To our knowledge, only one systematic review focused on the CPM reliability and analyzed the data qualitatively [7]. The review included 10 studies that investigated both healthy participants and patients, nine of which reported relative measures, one reported an absolute measure, and five studies reported both [24, 28–36]. After publication of that review, the number of additional studies published on CPM reliability have nearly doubled (16 studies), suggesting that an updated evidence synthesis is warranted. Performing a meta-analysis will provide quantitative information on CPM relative and absolute reliability measures, which has not been provided before [37].

Reliability can be measured in different ways. We focus on relative and absolute reliability for intra-session (test repeated in the same day) and inter-session (test repeated on different days) measures. The intraclass correlation coefficient (ICC) is the most common relative reliability measure used. It describes the resemblance between results in the same group, i.e., the degree of similarity between different ratings collected from the group [38, 39]. The absolute reliability is the degree of difference between different ratings collected from the same group, mostly reported by Bland Altman limit of agreement (LoA) [40]. The current systematic review and meta-analysis aimed to evaluate the intra-session and inter-session reliability of CPM in both healthy individuals and chronic pain patients. In addition, we analyzed the reliability according to gender and age.

Methods

We followed the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) 2018 guideline for systematic reviews and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [41] reporting guidelines (PROSPERO registration no.: CRD42021236300). Neither patients nor healthy individuals were involved in the design or planning of this study.

Study selection

Relevant articles were selected and included if they met the following criteria:

- Participants: healthy individuals or patients with chronic pain.
- Design: repeated measures of pain modulation, specifically diffuse noxious inhibitory control (DNIC, old term later replaced by CPM) or CPM, in two or more identical sessions, and studies that reported the DNIC or CPM reliability.
- Measurement procedure: measurements performed using one or multiple TS and CS to induce CPM effect.
- Outcome: estimates of CPM relative reliability (ICC) ranging from 0 to 1, and absolute reliability (LoA) were reported.

There were no specific restrictions regarding date of publication, language, age, race, number of participants, intervention, CPM assessment methods, or whether CPM was a primary or secondary outcome. Literature reviews, abstracts, editorial commentaries, letters to the editor, animal studies, and studies that did not assess DNIC/CPM, or assessed CPM only once were excluded.

Search strategy

A comprehensive search of electronic databases was performed from inception to April 2021, including EMBASE (OVID), Web of Science (Thomson Reuters), and NCBI (Pubmed). Broad keywords were created with the help of a librarian, and used to meet our aim: (1) Pain Modulation, (2) Conditioned Pain Modulation, (3) Diffuse Noxious Inhibitory Control, and (4) Dynamic sensory testing. Besides, reference lists of articles were hand-searched for relevant studies, and MeSH terms were used to search NCBI (PubMed) (Appendix A). Titles and abstracts were screened independently. One researcher (RN) selected studies for inclusion in a two-step process. First, studies titles and abstracts were screened. Second, full-text studies were reviewed to identify eligibility papers.

Outcomes and collected data

The following data were extracted from each article according to availability:

- Participants: gender, age, and health status.
- Study Design: sample size, number of visits (defined as testing procedures on different days), number of sessions (defined as testing procedures in the same day), the interval between sessions or visits, type of test and conditioning stimuli, and tested body site.
- Outcome: reliability measures for baseline TS and CPM effect, ICC estimates as well as the LoA estimates. The reliability of the TS was assessed to complement the analysis of CPM reliability and contribute to the explanation of the CPM findings. The reliability of CPM may be influenced by the reliability of the TS, because the CPM measurement is based on changes in the response to this stimulus.

The ICC estimates were stratified individually for each study. For example: Cathcart et al. reported two ICC estimates, one for each body site [29]. The ICC findings were stratified as: shoulder 0.69, finger 0.57. ICC estimates were classified as poor when <0.4, fair if 0.4–0.59, good if 0.6–0.75, and excellent when >0.75 [39]. All estimates were identified as intra-session (tests within the same day) and inter-session (tests on different days). Estimates for TS and CPM reported were grouped into four ICC classes (excellent, good, fair, and poor) for inter-session and intra-session reliability in healthy individuals and patients.

Risk of bias assessment

Consensus-based standards for the selection of health measurement instruments – risk of bias checklist (COSMIN-ROB): The COSMIN-ROB tool assesses the methodological quality of studies on reliability and measurement error of outcome measurement instruments. We used the extended version of the COSMIN-ROB for PROM. The (COSMIN-ROB) checklist (Consensus-based standards for the selection of health measurement instruments) assesses the risk of bias. Each study was assessed on a 4-point scale as 'very good', 'adequate', 'doubtful' or 'inadequate'; the overall rating of its quality was determined by the lowest rating (i.e., "the worst score counts" principle) [42]. Next, the results of each study were rated against the updated criteria for good measurement properties. Specifically, each result was rated as either sufficient (+) if ICC or weighted Kappa ≥0.70, insufficient (–) if ICC or weighted Kappa ≤0.70, or indeterminate (?) if ICC or weighted Kappa was not reported (Appendix B Table S1).

Quality appraisal for clinical measurement research reports evaluation form (QACMRR): The overall quality of individual studies was appraised using a 12-items structured clinical measurement-specific appraisal tool [43]. The evaluation criteria were: (1) thorough literature review to define the research question; (2) specific inclusion/exclusion criteria; (3) specific hypotheses; (4) appropriate scope of psychometric properties; (5) sample size; (6) follow-up; (7) the authors referenced specific procedures for administration, scoring and interpretation of procedures; (8) measurement techniques were standardized; (9) data were presented for each hypothesis; (10) appropriate statistics-point estimates; (11) appropriate statistical error estimates; and (12) valid conclusions and recommendations. An article's total score reflecting the quality of each study was presented as a percent: sum of scores for each item, divided by the numbers of items and multiplied by 100%. The quality summary of appraised articles ranged from poor to excellent, where (0–30%) was poor, (31–50%) was fair, (51–70%) was good, (71–90%) was very good, and (>90%) was excellent [43].

Statistical analysis

All descriptive data including participants age, number of visits, number of sessions, interval between tests, type of stimuli used (TS and CS), ICC estimates, and checklists findings were presented as range of values, mean ± standard deviation (SD), percent, or score.

Intraclass correlation coefficients (ICC) reported in the original articles were used for data analysis. ICC range from 0 to 1; approaching 1 represents strong reliability. An ICC less than 0.4 was considered poor agreement; 0.4–0.59 fair agreement; 0.6–0.75 good agreement; and greater than 0.75 excellent agreement [39].

A meta-analysis of reliability coefficients was performed in STATA (StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC) with meta package. The meta-analyses were conducted using a random effects model and the coefficients were converted to z values. Heterogeneity was deemed substantial if I^2 values were more than 50%. Forest plots were created using 95% CIs for coefficient estimates. Negative ICC estimates were excluded from the analysis, as negative estimates indicate that the true ICC is low, meaning that two members chosen randomly from any class can vary almost as much as any two randomly chosen members of the whole population [44]. In the presence of substantial statistical heterogeneity, univariate meta-regressions were performed to explain the sources of heterogeneity. Bubble plots were utilized with 95% confidence intervals. Regarding the Bland-Altman LoA analysis, the closer to zero the values, the lower the bias and higher the agreement. In this paper the LoA estimates were reported as ranges and means ± standard deviations (SD).

Results

Study selection

Our search yielded 40,406 articles. After removal of duplicates, title and abstract screening, and excluding non-CPM related papers, 656 papers were eligible for further review. Next, studies that assessed CPM once or did not report CPM reliability estimates were excluded, leaving 26 articles selected for full-text review. Of these, 25 papers were eligible (Figure 1) [23–36, 45–56]. Nine of Kennedy's studies [7] have been identified and included in our meta-analysis, one was excluded as it did not report relative reliability.

Study characteristics

Target population: Among all included records, 21 studies (84%) investigated healthy individuals (n=698, 33.3 ± 15.8),

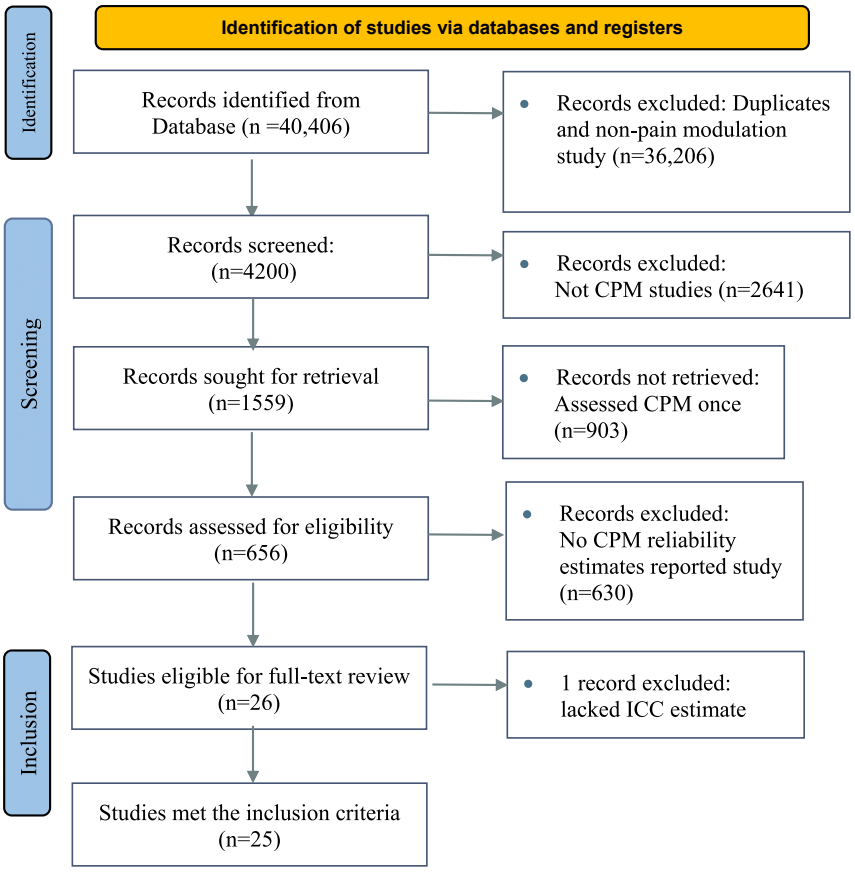


Figure 1: Study flow diagram of eligible studies [72]. ***CPM**, conditioned pain modulation.*

three (12%) chronic pain patients (two chronic back pain and one painful chronic pancreatitis n=206, 68.7 ± 18), and one study (4%) included both (n=324, healthy n=190 and musculoskeletal pain patients n=134) (Table 1).

Sample size: Across all studies, the sample size ranged from 20 to 342, mean 49.1 ± 60.4. The number of participants ranged from 20 to 26 individuals in 12 studies (48%), 30–36 in three (12%) studies, 60–65 in three (12%), and 50–55 in two (8%) studies. Two (8%) studies had 42 participants, one study (4%) 70, one study (4%) 89, and one study (4%) 342 for their 134 patients and for their 190 healthy individuals (Table 1).

Testing frequency: Two studies (8%) performed their tests in one visit; one had two sessions, and one had three sessions. One study (4%) included ranged from 1 to 8 visits. Eighteen studies (72%) tested their individuals in two visits, with each consisting of one session, except for one study; their first visit had two sessions and their second visit had one session. Two studies (8%) had three visits, one of which had one session and one had two sessions per visit. Two studies (8%) examined their individuals in four visits with two sessions each, one study (4%) had eight visits with one session each, and one (4%) study tested their heathy individuals group in three visits and their patients group in two visits each of which had one session per visit (Table 2). The interval time between visits ranged from one day to four months. One of the two single visit studies reported 60 min interval between two sessions, and one study did not specify (Table 2).

Test stimulus (TS): Nineteen (76%) studies used one type of TS, five (20%) two types, and one (4%) four types. Pressure stimulus mainly using the pressure pain threshold (PPT) was the most common TS, used in fifteen (60%) studies, followed by heat in ten (40%), electrical in four (16%), and ischemic pressure (IP) in four studies (16%) (Table 2).

Conditioning stimulus (CS): Nineteen (76%) studies used one type of CS, five (20%) used two types, and one (4%) used three types. Cold pressor test (CPT) test was the most common CS, used in 16 studies (64%), IP in six (24%), hot water bath (HWB) in four (16%), heat in two (8%), and pressure in two (8%). Two studies (8%) used control testing, where lukewarm water was used as a CS in comparison with HWB, heat or CPT (Table 2).

Intensity of the conditioning stimulus: Regarding the CPT intensity, the temperature used ranged from 0 to 12 °C, where one (4%) study used 0 °C, four (16%) 2 °C, two (8%) 4 °C, one (4%) 5.2 °C, two (8%) 7 °C, one (4%) 8 °C,

Table 1: Eligible studies: participants' characteristics.

Study	Year	n	Male	Healthy	Patient	Female	Healthy	Patient	Age mean ± SD		
									Overall	Male	Female
1. Cathcart et al.	2009	20	9	9	–	11	11	–	–	27 ± 6.4	23 ± 3.6
2. Olesen et al.	2012	62	38	–	38	24	–	24	53 ± 11.0	–	–
3. Lewis et al.	2012	20	7	7	–	13	13	–	25 ± 8.0	–	–
4. Martel et al.	2013	55	20	–	20	35	–	35	–	48.9 ± 10.5	49.5 ± 8.9
5. Valencia et al.	2013	324	161	74	87	163	116	47	H: 23.02 ± 6.04	–	–
		H: 190							P: 43.83 ± 17.80		
		P: 134									
6. Wilson et al.	2013	22	–	–	–	22	22	–	–	–	27 ± 7.0
7. Biurrun-Manresa et al.	2014	34	34	34	–	–	–	–	–	27.5 ± 6.8	–
8. Jurth et al.	2014	*36	–	–	–	–	–	–	–	–	–
9. Vuilleumier et al.	2015	89	39	–	39	50	–	50	56 ± 15.9	–	–
10. Imai et al.	2016	26	26	26	–	–	–	–	–	25.3 ± 5.7	–
11. Gehlling et al.	2016	25	13	13	–	12	12	–	31.6 ± 14.3	–	–
12. Granovsky et al.	2016	65	15	15	–	20	20	–	26.1 ± 2.5	–	–
		Study 1: 35	15	15	–	15	15	–	25.9 ± 2.6	–	–
		Study 2: 30									
13. Bossmann et al.	2016	20	9	9	–	11	11	–	34.4 ± 10.2	–	–
14. Costa et al.	2017	20	–	–	–	20	20	–	–	–	21.6 ± 1.81
15. Marcuzzi et al.	2017	42	21	21	–	21	21	–	30.2 ± 10	–	–
16. Graven-Nielsen et al.	2017	20	10	10	–	10	10	–	30 ± 5	–	–
17. Lie et al.	2017	24	14	14	–	10	10	–	25.8 ± 3.8	–	–
18. Lie et al.	2019	25	14	14	–	11	11	–	24.1 ± 3.7	–	–
19. Kennedy et al.	2019	50=	24=	A: 9	–	26=	A: 7	–	Total: 33.9 ± 11.27	–	–
		A: 16	A: 9	B: 7		A: 7	B: 10		A: 29.7 ± 7.98		
		B: 17	B: 7	C: 8		B: 10	C: 9		B: 32.1 ± 7.70		
		C: 17	C: 8			C: 9			C: 39.8 ± 14.56		
20. Larsen et al.	2019	22	15	15	–	7	7	–	23.6 (±2.4)	–	–
21. Alsouhibani et al.	2019	30	15	15	–	15	15	–	19.3 ± 1.5	–	–
22. Hoegh et al.	2019	25	25	25	–	–	–	–	–	30.2 ± 10.8	–
23. Nuwailati et al.	2020	60	22	22	–	38	38	–	–	34.2 ± 14.7	37.6 ± 15.1
24. Naugle et al.	2020	70=	37=	21 old	–	33=	19 old	–	67.6 ± 64.8 old	–	–
		40 old	21 old	16 young		19 old	14 young		22.8 ± 63 young		
		30 young	16 young			14 young					
25. Cummins et al.	2020	42=	19=	13	–	23=	12	–	–	32 ± 11	25 ± 9
		25 thermal	13	6		12	11			35.83 ± 11.23	31.82 ± 4.09
		17 mechanical (IP)	6			11					

n, sample size; **H**, healthy; **P**, patient. **Of 40 volunteers included in this study (20M and 20F), four were excluded from analysis. The number of participants of each gender after exclusion was not reported. Data reported as published in the original paper (age presented as mean ± SD). Otherwise, when possible, numbers were calculated for standardized presentation.*

Table 2: Eligible studies: methodological description.

Study	Visits	Sessions	Interval between sessions			Test stimulus, TS		Conditioning stimulus, CS		Reliability measures
			Min.	Days	Mon.	Type	Site	Type/intensity	Site	
1. Cathcart et al.	1	2	60	–	–	– Pressure: PPT	– Rt.M. finger – Rt.U. shoulder	– IP: 20 mmHg/s – VAS 3/10	– Lt.U. arm	ICC, CR
2. Olesen et al.	2	1	–	7	–	– Pressure: PPTol	– D. knee	– CPT: 2 °C	– Rt. hand	ICC
3. Lewis et al.	2	3	15	3 ± 0.5	–	– Pressure: PPT	– Rt. knee	– IP: 240 mmHg – CPT 12 ± 1 °C	– Lt. U. arm – Lt. hand	ICC
4. Martel et al.	2	1	–	7–10	–	– Pressure: PPT	– Rt. shoulder (U. trapezius)	– CPT: 4 °C	– Lt. hand	ICC, ISC
5. Valencia et al.	P: 2 H: 3	1	–	H: 3 and 5	P: 3	– Heat: SHPR	– NS. hand – ND. hand	– CPT: 8 °C	– S. hand – D. hand	ICC, SEM, MDC
6. Wilson et al.	8	1	–	18–24	–	– Heat: VNPS	– D. forearm	– HWB: 46.5 °C	– ND. hand	ICC
7. Biurrun-Manresa et al.	2	1	–	7–21	–	– Electrical: NWR	– D. thigh	– CPT: <2 ± 0.1 °C	– ND. hand	ICC, CV, B-A LoA
8. Jurth et al.	2	1	–	28	–	– Electrical: NFR	– Lt. thigh	– HWB: 46.5 °C – Control water: 33 °C	– Rt. hand	ICC
9. Vuilleumier et al.	2	1	–	10.0 ± 4.6	–	– Pressure: PDT	– Foot	– CPT: <2 °C	– NP. hand	ICC, CV, CR, B-A LoA
10. Imai et al.	2	1	–	8.6 ± 2.6	–	– Electrical: EPT – Heat: HPT – Pressure: PPT – IP – PDT PTTol	– D. ankle – D.M.V. forearm – Lateral of D.L. leg back of D.L. leg	– CPT: 0–4 °C – IP: 1 kPa/s – VAS 7/10	– ND. hand – ND.L. leg	ICC, CV
11. Gehlling et al.	2	1	40 ± 19.9	–	–	– Heat: NRS	– Rt.V. forearm	– CPT: 10 °C	– Lt. hand	ICC, SRD, SEM, B-A LoA
12. Granovsky et al.	2	1	–	Study 1: 3–7 Study 2: 7 days	–	– Heat: NPS	– ND.V. hand – ND.V. forearm	– HWB 46 °C – Heat: 0.5 °C above the TS temperature which was at baseline of 32 °C, rate of 4 °C to reach 30/100 NPS.	– D. hand – D.U. arm	ICC
13. Bossmann et al.	2	1	–	3	–	– Pressure: PPT	– D. hand	– Inter-digital web pinching: VAS 4–6/10	– ND. and	ICC, CV, SEM
14. Costa et al.	2	1	–	7	–	– Pressure: PPT	– D. side of the face	– HWB: 45.5–46.5 °C	– ND. hand	ICC, CV
15. Marcuzzi et al.	3	1	–	2 mon.: 58.6 ± 8.8 days 4 mon.: 116 ± 11.8 days	–	– Heat: NRS – Pressure: PPT	– ND.V. forearm – ND.L. Neck (U. trapezius)	– CPT: 10.5 ± 1 °C	– D. foot	ICC, SEM, MDD

Table 2: (continued)

Study	Visits	Sessions	Interval between sessions			Test stimulus, TS		Conditioning stimulus, CS		Reliability measures
			Min.	Days	Mon.	Type	Site	Type/intensity	Site	
16. Graven-Nielsen et al.	2	1	–	–	1	– IP: PDT – PVAS PTTol – Pressure: PPT	– U. arm – L. leg – Thigh	– IP: 10,30,60 kPa – VAS 7/10	Contra-lateral: – U. arm – L. Leg	ICC, CV, B-A LoA
17. Lie et al.	2	1	–	14.2 ± 6.8	–	– Heat: VAS – Tonic & phasic	– V. forearm	– CPT: 7 °C	Contra-lateral: – Hand	ICC, CV, B-A LoA
18. Lie et al.	2	1	–	7	–	– Heat: VAS – Electrical: NWR	– D.V. forearm – D.L. leg (tibialis anterior)	– CPT: 7 °C	– ND. hand	ICC, B-A LoA, bias p-value
19. Kennedy et al.	2	1	15	28.2 ± 4.8	–	– Pressure: PPT	– Rt. forearm	– Heat: 46.5 °C – CPT: 12 °C – Sham: 24 °C	– Lt.V. forearm – Lt. hand – Lt. hand	ICC, SEM
20. Larsen et al.	2	1	–	1–2	–	– Pressure: NRS	– D.L. leg	– Pressure: 1.3 kg force	– D. earlobe	ICC, SEM, B-A LoA
21. Alsouhibani et al.	2	1	23	7	–	– Pressure: PPT	– Rt. shoulder – Rt. thigh	– CPT: 0 ± 1 °C	– Lt. foot	ICC
22. Hoegh et al.	1	3	–	–	–	– IP: PTTol	– L. leg	– IP: PTTol 1 kPa/s, 70% of the subjects' PTTol as the conditioning pressure.	– D.L. leg	ICC
23. Nuwailati et al.	3	6	–	1st and 2nd 3.5 ± 1.4 days 2nd and 3rd 3.1 ± 2.8	–	– Pressure: PPT	– D. side of the face – D. hand – D. foot	– CPT: 5 °C	– ND. hand	ICC, CV, B-A LoA, SEM, bias p-value
24. Naugle et al.	2	1	–	2	–	– Pressure: PPT – Heat: VAS	– Lt. ventral forearm	– CPT: 10 °C	– Rt. hand	ICC, ISC
25. Cummins et al.	4	2	–	67 ± 63 28 ± 12	–	– Heat (HPT) – IP: PTTol – PDT	– Lt.V. forearm – D.L. leg	– CPT: 4 °C – IP: 1 kPa/s, 70% of the subjects' PTTol as the conditioning pressure.	– ND. Rt. foot – ND. Rt. forearm – ND.L. leg	SEM, ICC

min, minutes; **mon**, months; **H**, healthy; **P**, patients; **Lt**, left; **Rt**, right; **U**, upper; **L**, lower; **M**, middle; **D**, dominant; **ND**, none dominant; **S**, surgical; **NS**, none surgical; **NP**, none painful; **V**, volar; **Caud**, caudal; **IP**, ischemic pressure; **PPT**, pressure pain threshold; **PTTol**, pain tolerance threshold; **CPT**, cold pressor test; **MP**, mechanical pressure; **HWB**, hot water bath; **SHPR**, supra-threshold heat pain response; **VNPS**, verbal numerical pain scale; **NRS**, numerical rating scale; **NPS**, numerical pain scale; **PVAS**, pressure value when electronic visual analogue scale was 6; **CoVAS**, computerized visual analog scale; **HPT**, heat pain threshold; **EPT**, electric pain threshold; **EPD**, electric pain detection; **PIR**, pain intensity rating; **IDW**, Inter-digital web; **NFR**, nociceptive flexion reflex; **NWR**, nociceptive withdrawal reflex; **PDT**, pain detection threshold; **ICC**, intra-class correlation; **CR**, coefficient of repeatability; **CV**, coefficient of variation; **ISC**, intra-individual stability coefficient; **SEM**, standard error of measurements; **MDC**, minimum detectable change; **SRD**, smallest real difference; **MDD**, minimum detectable difference; **B-A LoA**, Bland-Altman limits of agreement.

three (12%) 10 °C, and two (8%) 12 °C. The HWB stimulus intensity was reported in four (16%) as 45–46.5 °C. One study (4%) reported their heat intensity as 46.5 °C. The IP stimulus was reported by one study (4%) as 20 mmHg/s with VAS score of 3/10, and one (4%) as 240 mmHg. Two pressure (8%), one heat (4%), and four IP (16%) intensities were subjective; the intensity of the CS was determined according to the TS intensity used for each individual prior to their CPM testing (Table 2).

Tested sites: The TS was applied on the forearm in 10 studies (40%), the lower leg in six (24%), the hand in five (20%), the thigh in four (16%), the shoulder in three (12%), the side of the face in two (8%), the foot in two (8%), the knee in two (8%), the ankle in one (4%), the upper arm in one (4%), and the neck in one (4%) study (Table 2).

The CS was applied to the hand in 18 studies (72%), the upper arm in four (16%), the lower leg in four (16%), the foot in three (12%), the forearm in two (8%), and the earlobe in one (4%) (Table 2).

Reliability

Estimates for TS and CPM reported below were grouped into four ICC classes (excellent, good, fair, and poor) for inter-session and intra-session reliability in healthy individuals and patients.

Test stimulus (TS) reliability

A total of 19 studies reported 85 ICC estimates for TS reliability, which ranged from 0.15 to 0.96.

***Class of ICC*:** The majority were excellent ICC 63 (74%), 12 (14%) were good, six (7%) were fair, and four estimates were poor (5%).

***Testing Session*:** 71 (83%) were intra-session estimates and 14 (17%) were inter-session.

***Target population*:** 76 (89%) estimates were reported from healthy subjects, while nine (11%) were from patients.

***Meta-analysis*:** The **intra-session** reliability was excellent when electric, IP and pressure were applied on healthy subjects: electric (k=3, 6 estimates, 163 subjects, ICC=0.93, 95% CI: 0.89–0.95, I^2=45.52%), IP (k=2, 22 estimates, 488 subjects, ICC=0.88, 95% CI: 0.85–0.91, I^2=39.36%), and pressure (k=10, 21 estimates, 628 subjects, ICC=0.83, 95% CI: 0.78–0.87, I^2=65.77%).

The ICC was good when heat was applied to healthy individuals (k=7, 19 estimates, 586 subjects, ICC=0.65, 95% CI: 0.53–0.74, I^2=77.59%), and when pressure was applied to patients (k=3; 3 estimates, 206 patients, ICC=0.64, 95% CI: 0.44–0.78, I^2=76.21%) (Figure 2).

A univariate meta-regression was conducted to investigate the sources of the moderate statistical heterogeneity for pressure TS intra-session reliability in healthy subjects. The regression coefficient for the field of testing sites was statistically significant with p<0.001 (–0.19, 95% CI: –0.27, –0.11), with 67% of between-study variance explained (R^2=66.95%) (Figure 6).

The **inter-session** reliability was excellent when pressure was applied to both heathy individuals (k=4, 8 estimates, 340 subjects, ICC=0.86, 95% CI: 0.78–0.91, I^2=78.64%), and patients (k=1, 6 estimates, 534 patients, ICC=0.91, 95% CI: 0.89–0.93, I^2=40.02%) (Figure 3). No inter-session estimates were reported from healthy individuals or patients for other test stimuli.

***LoA*:** Six studies reported 31 estimates from 892 healthy subjects, ranging from –142 to 11, overall mean –13 ± 30.1. All six studies reported 28 intra-session estimates (–13.9 ± 31.66) and one study reported three inter-session estimates (–8.3 ± 6.8).

A univariate meta-regression was conducted to investigate the sources of the moderate statistical heterogeneity for pressure TS intra-session reliability in healthy subjects. The regression coefficient for the field of testing sites was statistically significant with p<0.007, (–0.18, 95% CI: –0.31, 0.05), with 98% of between-study variance explained (R^2=60.91%) (Figure 6).

Conditioned pain modulation (CPM) reliability

A total of 25 studies reported 94 ICC estimates for CPM reliability, which ranged from –0.45 to 0.94. Meta-analysis was performed on 88 estimates.

***Class of ICC*:** 10 (11%) estimates were excellent, 16 (18%) good, 28 (32%) fair, and 34 poor (39%).

***Testing Session*:** 24 (27%) were intra-session estimates and 64 (73%) were inter-session.

***Target population*:** 81 (92%) reported from healthy subjects, while seven (8%) estimates only were reported from patients.

Meta-analysis: The **intra-session** reliability was good when pressure and IP were applied to healthy subjects: pressure as TS with CPT or IP as CS (k=4, 11 estimates, 360 subjects, ICC=0.64, 95% CI: 0.45–0.77, I^2=81.74%), and IP as TS and CS simultaneously (k=2, 3 estimates, 75 subjects, ICC=0.62, 95% CI: –0.17 to 0.92, I^2=92.64%,). The intra-session reliability was fair when heat was applied as TS to healthy individuals, with CPT as CS (k=2, 6 estimates, 645 subjects, ICC=0.55, 95% CI: 0.44–0.64, I^2=57.14%).

When heat was applied to patients as TS with CPT as CS, the reliability was poor (k=1, 2 estimates, 268 patients,

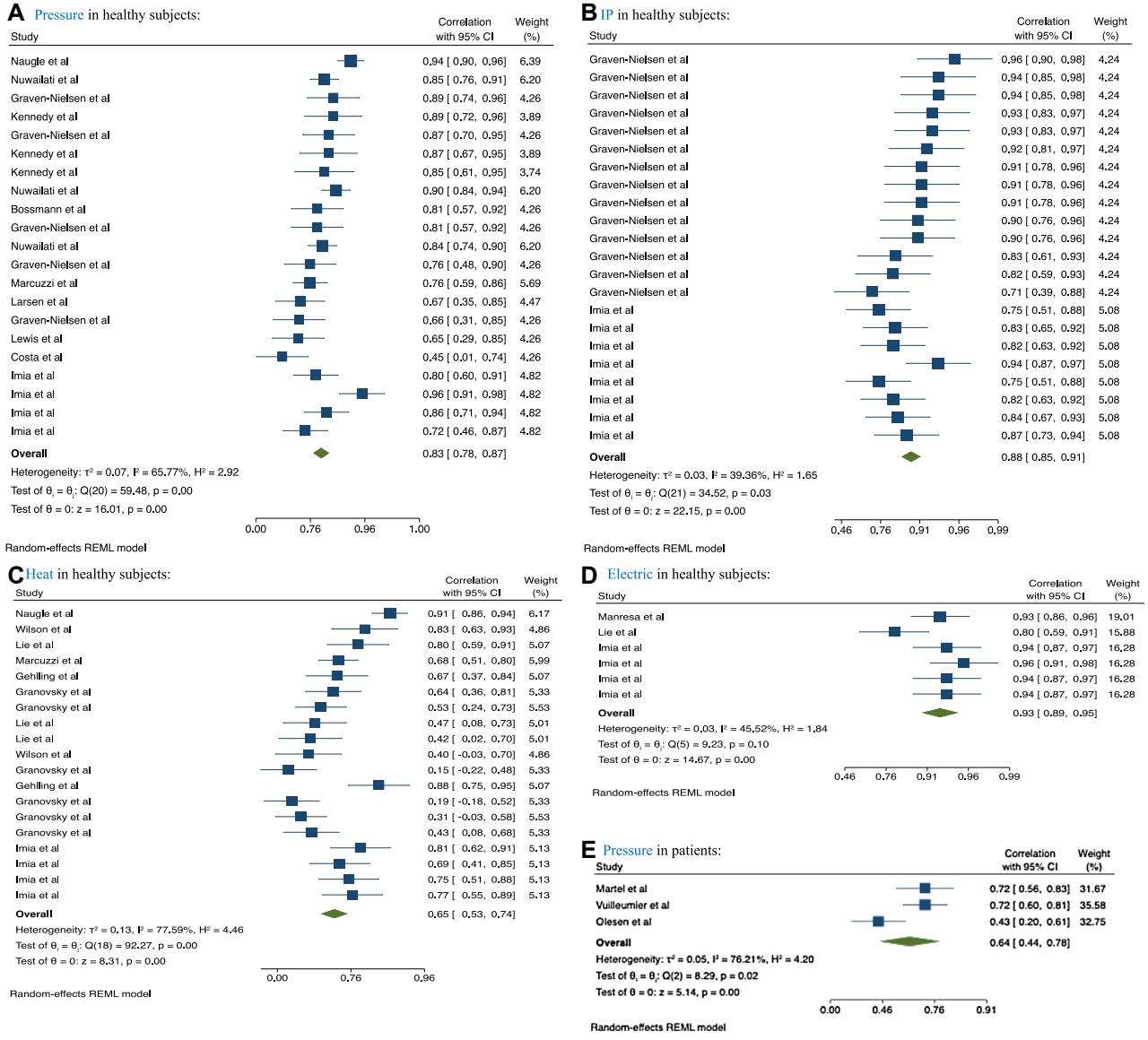


Figure 2: Forest plot for meta-analysis of TS intra-session reliability: **A. Pressure** TS in healthy subjects, **B. IP** TS in healthy subjects, **C. Heat** TS in healthy subjects, **D. Electric** TS in healthy subjects, and **E. Pressure** TS in patient. ***ICC,*** Intraclass Correlation; ***IP,*** ischemic pressure; ***TS,*** test stimulus.

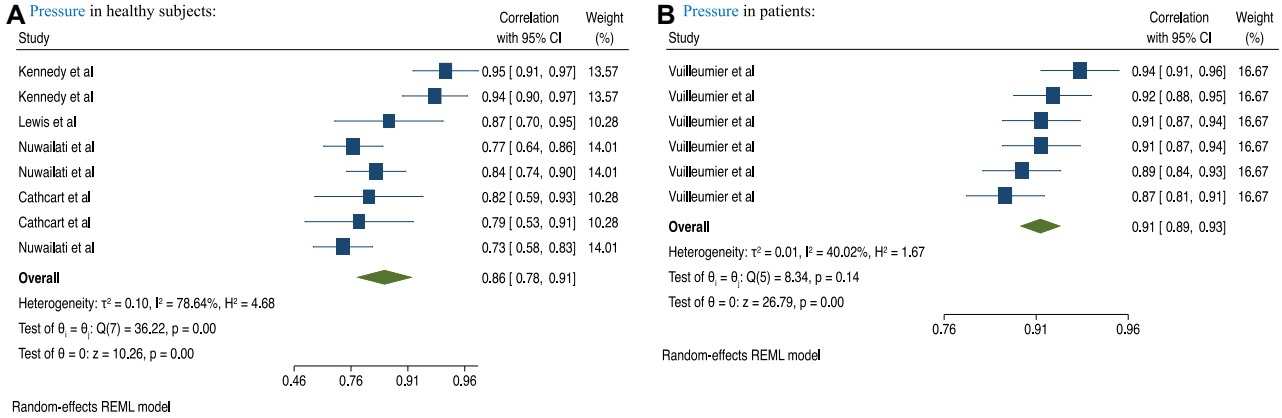


Figure 3: Forest plot for meta-analysis of TS inter-session reliability: **A. Pressure** TS in healthy subjects, **B. Pressure** TS in patients. ***ICC,*** Intraclass Correlation; ***TS,*** test stimulus.

ICC=0.25, 95% CI: −0.11 to 0.56, I^2=89.29%). However, pressure demonstrated excellent reliability as TS, with CPT as CS (k=1, 2 estimates, 178 patients, ICC=0.77, 95% CI: 0.70–0.82, I^2=0.00%) (Figure 4).

The **inter-session** reliability was fair when IP, pressure and heat were applied to healthy subjects:

IP as TS with IP or CPT as CS (k=2, 18 estimates, 384 subjects, ICC=0.51, 95% CI: 0.42–0.59, I^2=5.32%); pressure as TS with CPT, heat, IP, HWB, or pressure as CS (k=11, 21 estimates, 639 subjects, ICC=0.43, 95% CI: 0.30–0.54, I^2=66.93%); and heat as TS with CPT, heat or IP as CS (k=8, 14 estimates, 759 subjects, ICC=0.43, 95% CI: 0.32–0.53, I^2=60.19%).

The reliability of electric tests as TS was poor in healthy individuals with CPT, IP and HWB as CS (k=4, 8 estimates, 245 subjects ICC=0.35, 95% CI: 0.18–0.50, I^2=46.28%).

In patients, the inter-session reliability for pressure as TS with CPT as CS was fair (k=3; 3 estimates, 206 patients, ICC=0.44, 95% CI: 0.11–0.69, I^2=84.98%) (Figure 5).

LoA: Six studies reported 32 estimates from 917 healthy subjects, ranging from −18.1 to 42, overall mean 4.9 ± 11.8. One study reported three intra-session estimates with a mean of 5.3 ± 4.7, and six studies reported 29 estimates with a mean of 5.05 ± 12.6.

Gender and age

Gender: Twenty studies (80%) included both genders, two (8%) included men, two (8%) women, and one (4%) did not specify. Three studies (12%) reported ICC for both genders (Table 1). Collectively, the results suggest that CPM is more reliable in women than men. A subgroup analysis of Martel et al. found the test's reliability to be significantly higher in women (excellent ICC=0.75) than men (poor ICC=0.33) [28]. Overall, Nuwailati et al. also found the test to be slightly more reliable in women (ICC=0.27) than men (ICC=0.15), but for both genders the reliability was poor [45].

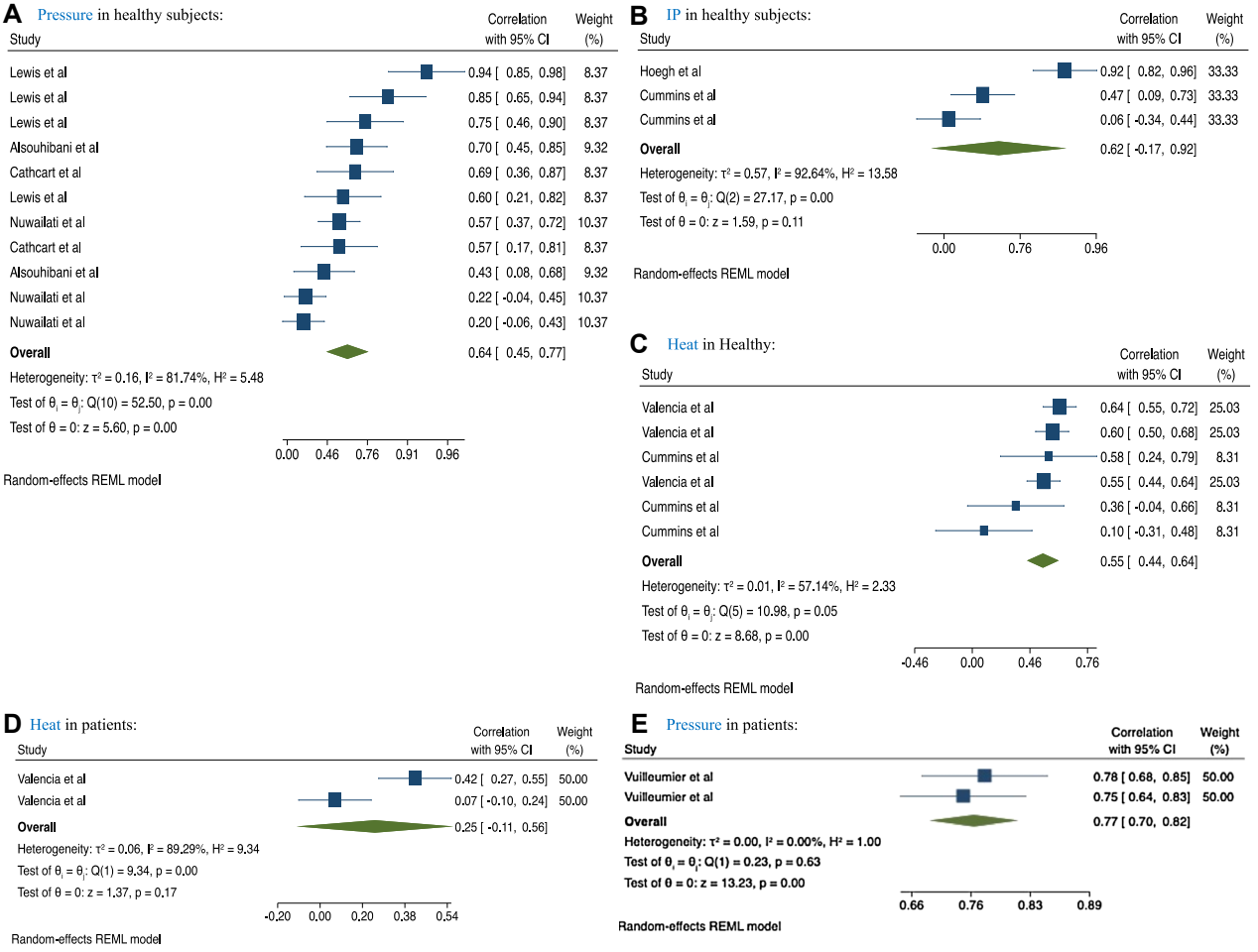


Figure 4: Forest plot for meta-analysis of CPM intra-session reliability: **A. Pressure** TS in healthy subjects, **B. IP** TS in healthy subjects, **C. Heat** TS in healthy subjects **D. Heat** TS in patients, **E. Pressure** in patients. ***ICC,*** Intraclass correlation; ***IP,*** ischemic pressure; ***TS,*** test stimulus; ***CPM,*** conditioned pain modulation.

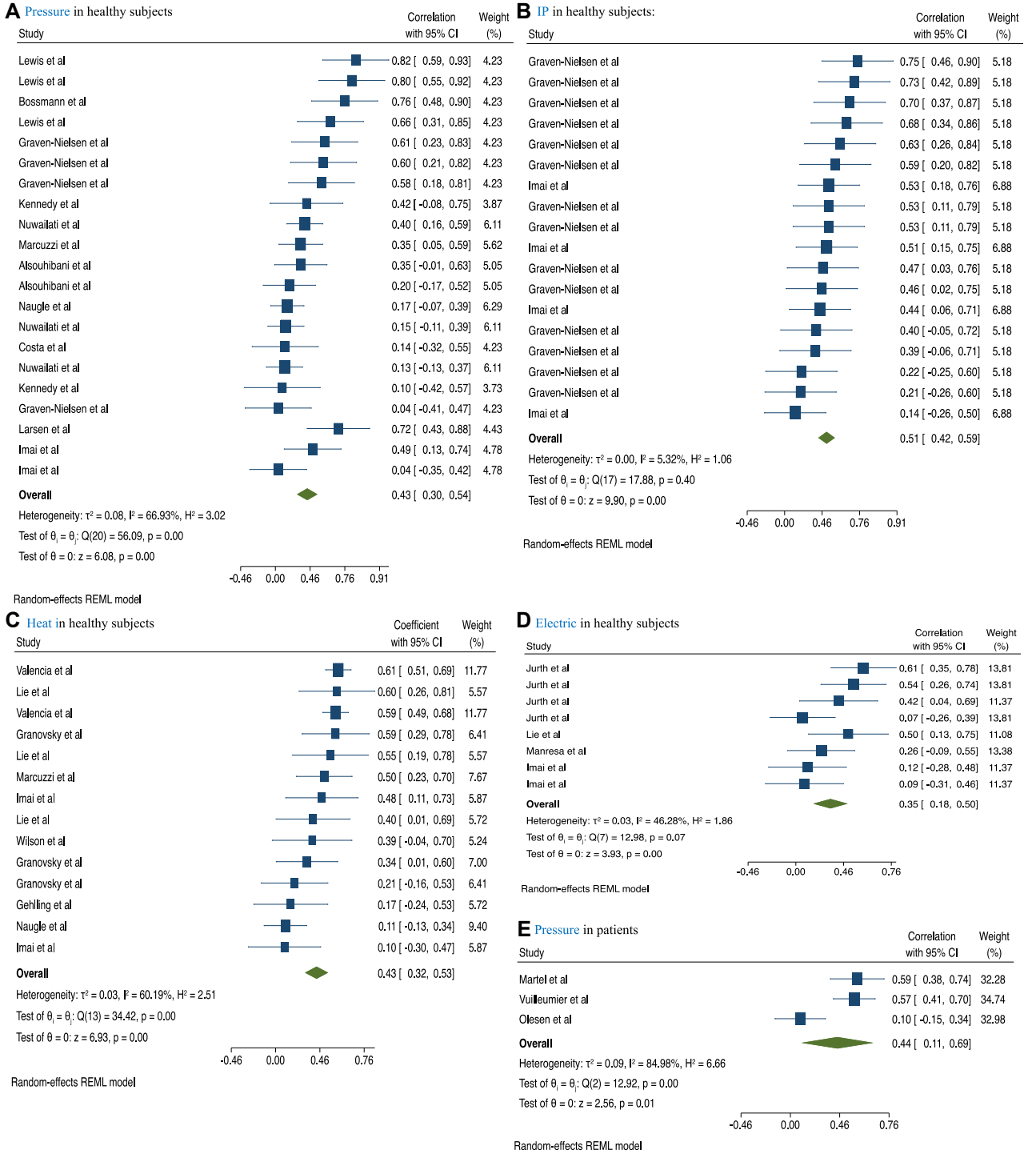
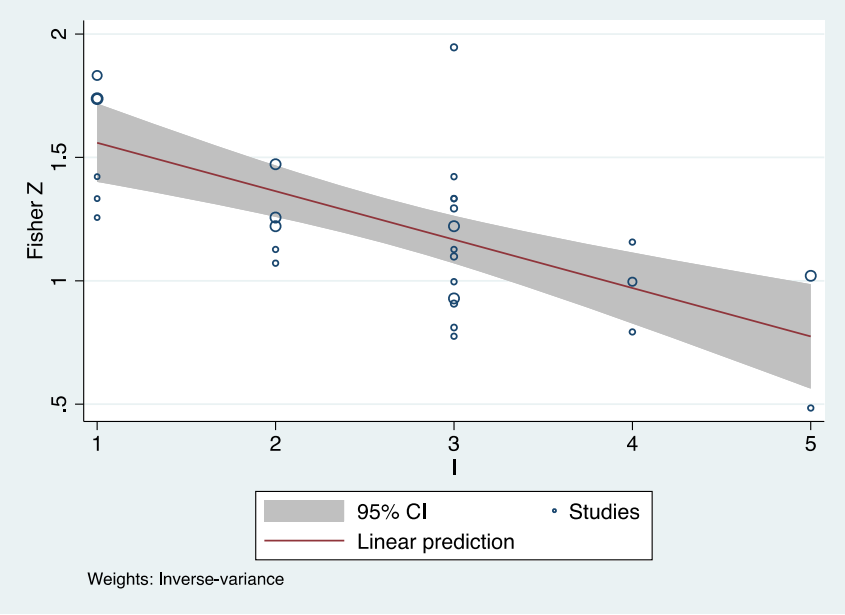


Figure 5: Forest plot for meta-analysis of CPM inter-session reliability: **A. Pressure** TS in healthy subjects, **B. IP** TS in healthy subjects, **C. Heat** TS in healthy subjects **D. Electric** TS in healthy subjects, and **E. Pressure** TS in patients. ***ICC,*** Intraclass correlation; ***IP,*** ischemic pressure; ***TS,*** test stimulus; ***CPM,*** conditioned pain modulation.

Valencia et al. found the test to be more reliable in women than men, both in healthy individuals and patients [32]. The intra-session reliability was good for female patients (ICC=0.65) and fair for men (ICC=0.40). In healthy individuals, the reliability was higher in women (ICC=0.63) than men (ICC=0.55). The inter-session reliability was very similar in healthy women (ICC=0.61) and men (ICC=0.59).

Age: Healthy participants' age ranged from 19.3 ± 1.5 to 67.6 ± 64.8 years, and patients' age from 43.83 ± 17.80 to 56 ± 15.9 years (Table 1). Two studies (8%) reported age



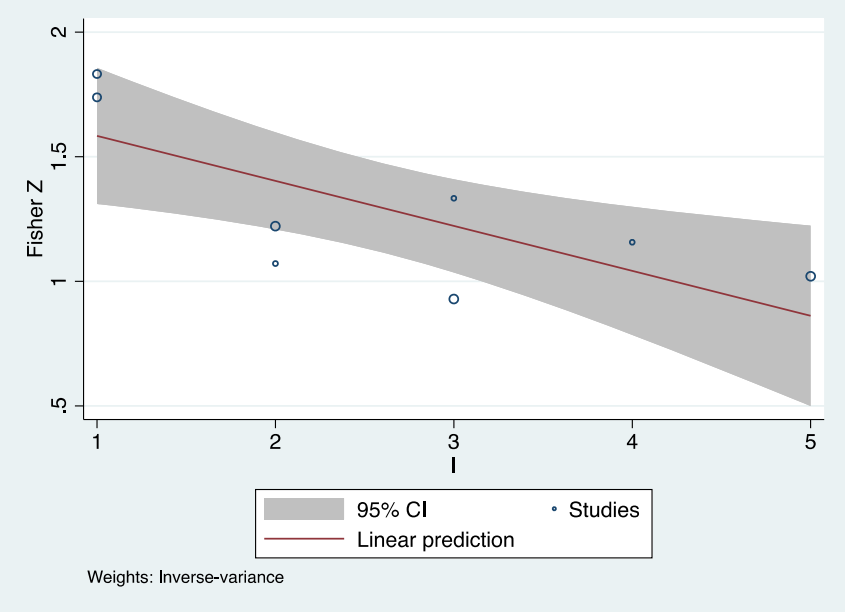


Figure 6: Bubble plot of pressure TS in healthy subjects. Testing sites: 1: forearm, 2: hand, 3: leg, 4: upper extremity, 5: face. A) Intra-session reliability B) inter-session reliability.

as an influential factor on test reliability. Naugle et al. found the test to be more reliable in a younger group (ICC=0.54) compared with an older group (ICC=0.14) [56]. Although the test had poor reliability in both age groups, Nuwailati et al. reported the test to be more reliable in 31–65 years old (ICC=0.36) than in 18–30 years old individuals (ICC=0.14) [45].

Risk of bias assessment

COSMIN-ROB checklist

The checklist is presented in Table 3. Item-1 was rated as adequate in twenty-four (96%) studies and as very good in one (4%), items 2, 6 and 7 were rated as very good in all twenty-five (100%) studies, item 3 was rated as very good in nineteen (76%) studies and as adequate in six (24%), item 4 was rated as very good in two (8%) studies and doubtful in twenty-three (92%), items 5 and 8 were rated as doubtful in all twenty-five (100%) studies, and item 9 was rated as inadequate in twenty-four (96%) studies and as very good in one (4%). As mentioned previously, the worst score determined the overall rating of the study quality. Therefore, as item 9 (a Kappa coefficient question) was rated as the worst, the overall rating of the 24 studies was inadequate; none of these studies reported the Kappa coefficient. The overall rating of one study was very good. Furthermore, the individual study assessment against the updated

criteria for good measurement resulted in 18 (72%) studies rated as insufficient (–) and 7 (28%) as sufficient (+) (Table 3).

Quality appraisal for clinical measurement research reports evaluation form QACMRR

All studies obtained an overall positive rating, four were rated as excellent papers (quality ranged from 92 to 96%), and 21 were rated as very good (71–88%). Mainly, two shortcomings were identified: the sample size determination item was addressed by nine studies (36%) only, and reporting a statistical error measurement estimates (SEM) was reported by eight studies (32%) (Table 3).

Discussion

This meta-analysis showed highly variable CPM reliability across studies, ranging from poor to excellent. Random error due to small sample sizes is a possible explanation for the variability. Despite the intensive research conducted throughout the past two decades, and experts' recommendations to adopt uniform and standardized methodology [4], the testing method of CPM still lacks standardization, possibly contributing to variability in reliability. Also, the inconsistent reliability findings may be related to the variable measures used for reporting results, such as using different measures to report absolute reliability.

Multiple factors have been found to influence the CPM effect, some of them non-modifiable such as age and gender, and other ones modifiable such as the CPM testing methodology [2]. The inter-session reliability was worse than the intra-session reliability. The poor inter-session reliability may not necessarily be due only to limitations is the CPM methodology and random error, but also to variability over time in endogenous inhibition, which may be a dynamic process rather than a stable trait. This is an essential point that should be the object of future mechanistic research. For instance, longitudinal studies may assess CPM in association with brain imaging measures of endogenous pain modulation to test whether changes in CPM over time are associated with changes in brain endogenous modulatory processes [57–59].

Table 3: Risk of bias assessment.

Study	Psychometric properties reported	COSMIN-ROB	COSMIN rating (criteria)	QACMRR
1. Cathcart et al.	Reliability	Inadequate	–	Very good
2. Olesen et al.	Reliability	Inadequate	–	Very good
3. Lewis et al.	Reliability	Inadequate	+	Very good
4. Martel et al.	Reliability	Inadequate	–	Very good
5. Valencia et al.	Reliability	Inadequate	–	Very good
6. Wilson et al.	Reliability	Inadequate	–	Very good
7. Biurrun-Manresa et al.	Reliability	Inadequate	–	Very good
8. Jurth et al.	Reliability	Inadequate	–	Very good
9. Vuilleumier et al.	Reliability	Inadequate	+	Very good
10. Imai et al.	Reliability	Inadequate	–	Very good
11. Gehlling et al.	Reliability	Inadequate	–	Excellent
12. Granovsky et al.	Reliability	Inadequate	–	Very good
13. Bossmann et al.	Reliability	Inadequate	+	Excellent
14. Costa et al.	Reliability	Inadequate	–	Very good
15. Marcuzzi et al.	Reliability	Very good	–	Very good
16. Graven-Nielsen et al.	Reliability	Inadequate	+	Very good
17. Lie et al.	Reliability	Inadequate	–	Very good
18. Lie et al.	Reliability	Inadequate	–	Very good
19. Kennedy et al.	Reliability	Inadequate	–	Excellent
20. Larsen et al.	Reliability	Inadequate	+	Very good
21. Alsouhibani et al.	Reliability	Inadequate	+	Very good
22. Hoegh et al.	Reliability	Inadequate	+	Very good
23. Nuwailati et al.	Reliability	Inadequate	–	Excellent
24. Naugle et al.	Reliability	Inadequate	–	Very good
25. Cummins et al	Reliability	Inadequate	–	Very good

Criteria for good measurement properties: +, sufficient; –, insufficient; ?, indeterminate. **COSMIN-ROB**, Consensus-based standards for the selection of health measurement instruments-risk of bias tool to assess the quality of studies on reliability and measurement error of outcome measurement instrument [29]; **QACMRR**, quality appraisal for clinical measurement research reports evaluation form [28, 29].

Methodological aspects

Although only four studies included patients, their total number of participant was acceptable (n=396); one study included 190 patients. The CPT was the most frequently utilized CS. The test reliability ranged from poor to fair with temperatures ranging from 0 to 10 °C, while the reliability was good when the CPT test was set at 12 °C. This is consistent with Kennedy et al. suggestion regarding using temperature between 8 and 12 °C. It is sufficient to induce CPM effect, and it is well tolerated by participants, especially chronic pin patients [7]. Lower temperatures (more intense) could lead the participants to violate the testing protocol, for instance with an early withdrawal of their hand from ice cold water.

Few studies examined different tested sites for both TS and CS, while others examined a single site only. It seems that the distance between TS and CS sites may be an influential factor on the test reliability. Particularly when both stimuli were applied contra-laterally on the same site or one segment away from each other [23, 27, 29, 32, 48, 54], their reliability ranged from fair to excellent. On the other hand, when the TS and CS were applied at sites distant from each other, the reliability was generally poor to fair [24–27, 31, 34, 35, 46, 50, 52, 54]. For instance, Oono et al. reported their best reliability when both stimuli were applied to the hand and forearm and to the upper arm and forearm, whereas the tests were less reliable when applied at head and forearm, and head and lower leg [30]. Imia et al. and Graven-Nelson et al. found better reliability when both stimuli were applied to the lower leg [25, 27]. Furthermore, the test seemed to be less reliable when applied to the side of the face and hand [45, 50].

Most of the estimates of TS ICC were excellent, but this was not associated with CPM reliability. The choice of the TS and CS type may have an impact on the test reliability. The CPT was suggested to be one of the most efficient stimuli to induce CPM, in particular when combined with pressure as a TS [60, 61]. Imai et al. and Graven-Nielsen et al. assessed multiple combinations of TS and CS. Among all, the IP demonstrated the highest inter-session reliability when used as TS and as CS, or when combined with CPT as CS [25, 27]. Findings of studies that used one TS were in line with the results of Imai et al. and Graven-Nielsen et al. The IP as a TS demonstrated an excellent inter- and intra-session CPM reliability, as well as pressure used as a TS when combined with either CPT or IP as a CS [24, 46, 54]. Our meta-analysis confirmed that the pressure and IP stimuli yielded the most reliable CPM test when used as TS and CS, or when used with CPT as CS.

Reliability according to gender and age

The test's reliability was slightly better in women than men in most studies, while a few studies reported no difference between genders. In women, Wilson et al. reported poor CPM stability, i.e. CPM fluctuations, during the follicular and luteal phases of the menstrual cycle [33], but no data were reported regarding the ovulatory phase. The CPM effect was reported to be less efficient during the follicular and luteal phases when compared to the ovulatory phase [62, 63]. The gender variability in CPM reliability may be due to several factors. It may reflect gender differences in the central nervous system (CNS) processing of pain. This explanation has been proposed by others as well; women exhibit lower μ-opioid receptor binding activity than men during the activity of opioidergic systems [64, 65]. Another factor that might have caused the variability is expectation [66]. Bjorkedal et al. suggested that expectation can influence the CPM effect, as information on the CPM effect affected women, but not men [66].

One study reported the test to be more reliable in younger individuals [56]. Similar to gender influence on reliability, numerous studies concluded that the CPM effect is more efficient in younger individuals, as older adults expressed less inhibition of cold stimulation pain [2, 67–70].

Risk of bias assessment

The COSMIN-ROB tool assesses the methodological quality of studies on reliability and measurement error of all types of outcome measurement instruments. We used the extended version 1.0 of the COSMIN-ROB for PROM dated in December 2020 [42]. All studies were published prior to this date, the latest was published in November 2020 [26]. All studies scored very good and adequate in multiple items. However, 96% scored inadequate in item-9 only, meaning that the kappa calculation was not reported by most of the studies. Cohen's kappa coefficient is used to measure intra and inter-rater reliability, and considers the possibility of the agreement occurring by chance [71]. The kappa coefficient was reported by one study only [53].

Because the scoring system of this particular checklist considers the worse score as the overall score for each study, all studies not reporting kappa calculation were scored overall as inadequate.

An additional identified issue was blinding. Blinding of participants or examiners in clinical trials is a key methodological procedure, and a certain degree of bias due to non-

blindness should be expected [72]. The professionals' knowledge of scores or values of other measurements (item-4 and 5) was under-reported. Blinding was reported in two studies only. In the study by Bossman et al., raters were blinded to each other's findings [48]. Kennedy et al. reported various forms of blinding: raters blinded to the findings of other raters, raters blinded to their prior findings, raters blinded to clinical information related to the subject, and any cues that were not part of the study design [53].

All studies had an excellent or good overall rating for QACMRR, but two shortcomings were identified. First, the samples size calculation (item-5) was under-reported; it was addressed by nine studies only. An example to follow is Biurrun-Manresa et al.'s sample size calculation, which was based on two parameters: the detection of a significant CPM effect and the reliability of that effect over time [34]. Furthermore, an appropriate statistical error estimate, such as the standard error of measurement (SEM) (item-11), was reported in only eight studies. It indicates the amount of variability in the test that is caused by measurement error and its effect on individual results. This finding stresses the importance of reporting the absolute reliability measurement for the CPM test.

Strengths and limitations

To the best of our knowledge, this is the first meta-analysis that quantified the reliability of CPM and provided a review of its influential factors. We added several studies to the previously published systematic review [7], thereby improving the precision of the reliability estimates. We used two different quality appraisal tools that helped identify methodological limitations in the included investigations and provide information to improve the quality of future reliability studies.

The data did not allow estimation of the absolute reliability. An attempt to conduct a meta-analysis for CS was challenging due to insufficient data for further stratification of estimate by stimulus type and tested site (18 estimates were reported by seven studies). It created smaller subgroups, thereby inducing lower statistical power. It is unclear whether the results for healthy participants can be generalized to patients due to the small number of participants.

Conclusions

The reliability of CPM test is highly variable across studies. Intra-session reliability using pressure and CPT were the TS and CS most consistently associated with good to excellent values in healthy volunteers and patients. The inter-session reliability was fair or less for all modalities, both in healthy volunteers and patients. An important goal of future research is to determine whether poor inter-session reliability is due to dynamic changes in endogenous inhibition, rather than to methodological limitations. Future studies can improve their methodology by including sample size calculation, reporting measures of absolute reliability (Bland-Altman, SEM, and kappa), and including blindness in their design. The use of CPM in clinical research and potentially in clinical practice will depend on improvement in its reliability, as well as more convincing evidence of its clinical usefulness.

Research funding: Pavlos Bobos was supported by a Frederick Banting and Charles Best Canada Graduate Scholarship (CGS-D) CIHR Doctoral Award and the Arthritis Society Postdoctoral Fellowship Award.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.
Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

1. Nir RR, Yarnitsky D. Conditioned pain modulation. Curr Opin Support Palliat Care 2015;9:131–7.
2. Hermans L, Van Oosterwijck J, Goubert D, Goudman L, Crombez G, Calders P, et al. Inventory of personal factors influencing conditioned pain modulation in healthy people: a systematic literature review. Pain Pract 2016;16:758–69.
3. Yarnitsky D, Arendt-Nielsen L, Bouhassira D, Edwards RR, Fillingim RB, Granot M, et al. Recommendations on terminology and practice of psychophysical DNIC testing. Eur J Pain 2010;14:339.
4. Yarnitsky D, Bouhassira D, Drewes AM, Fillingim RB, Granot M, Hansson P, et al. Recommendations on practice of conditioned pain modulation (CPM) testing. Eur J Pain 2015;19:805–6.
5. Arendt-Nielsen L, Larsen JB, Rasmussen S, Krogh M, Borg L, Madeleine P. A novel clinical applicable bed-side tool for assessing conditioning pain modulation: proof-of-concept. Scand J Pain 2020;20:801–7.
6. Schliessbach J, Lütolf C, Streitberger K, Scaramozzino P, Arendt-Nielsen L, Curatolo M. Reference values of conditioned pain modulation. Scand J Pain 2019;19:279–86.

7. Kennedy DL, Kemp HI, Ridout D, Yarnitsky D, Rice ASC. Reliability of conditioned pain modulation: a systematic review. Pain 2016; 157:2410–9.
8. Arendt-Nielsen L, Andresen T, Malver LP, Oksche A, Mansikka H, Drewes AM. A double-blind, placebo-controlled study on the effect of buprenorphine and fentanyl on descending pain modulation: a human experimental study. Clin J Pain 2012;28: 623–7.
9. De Kooning M, Daenen L, Roussel N, Cras P, Buyl R, Ickmans K, et al. Endogenous pain inhibition is unrelated to autonomic responses in acute whiplash-associated disorders. J Rehabil Res Dev 2015;52:431–40.
10. Oono Y, Wang K, Svensson P, Arendt-Nielsen L. Conditioned pain modulation evoked by different intensities of mechanical stimuli applied to the craniofacial region in healthy men and women. J Orofac Pain 2011;25:364–75.
11. Granovsky Y, Yarnitsky D. Personalized pain medicine: the clinical value of psychophysical assessment of pain modulation profile. Rambam Maimonides Med J 2013;4:e0024.
12. Mertens MG, Hermans L, Crombez G, Goudman L, Calders P, Van Oosterwijck J, et al. Comparison of five conditioned pain modulation paradigms and influencing personal factors in healthy adults. Eur J Pain 2021;25:243–56.
13. Lewis GN, Rice DA, McNair PJ. Conditioned pain modulation in populations with chronic pain: a systematic review and meta-analysis. J Pain 2012;13:936–44.
14. Giannoni-Luza S, Pacheco-Barrios K, Cardenas-Rojas A, Mejia-Pando PF, Luna-Cuadros MA, Barouh JL, et al. Noninvasive motor cortex stimulation effects on quantitative sensory testing in healthy and chronic pain subjects: a systematic review and meta-analysis. Pain 2020;161:1955–75.
15. Neelapala YVR, Bhagat M, Frey-Law L. Conditioned pain modulation in chronic low back pain: a systematic review of literature. Clin J Pain 2020;36:135–41.
16. O'Brien AT, El-Hagrassy MM, Rafferty H, Sanchez P, Huerta R, Chaudhari S, et al. Impact of therapeutic interventions on pain intensity and endogenous pain modulation in knee osteoarthritis: a systematic review and meta-analysis. Pain Med 2019;20:1000–11.
17. Yarnitsky D, Crispel Y, Eisenberg E, Granovsky Y, Ben-Nun A, Sprecher E, et al. Prediction of chronic post-operative pain: pre-operative DNIC testing identifies patients at risk. Pain 2008;138: 22–8.
18. Dürsteler C, Salazar Y, Rodriguez U, Pelfort X, Verdié LP. Conditioned pain modulation predicts persistent pain after knee replacement surgery. Pain Rep 2021;6:e910.
19. Niesters M, Proto PL, Aarts L, Sarton EY, Drewes AM, Dahan A. Tapentadol potentiates descending pain inhibition in chronic pain patients with diabetic polyneuropathy. Br J Anaesth 2014; 113:148–56.
20. Yarnitsky D, Granot M, Nahman-Averbuch H, Khamaisi M, Granovsky Y. Conditioned pain modulation predicts duloxetine efficacy in painful diabetic neuropathy. Pain 2012;153:1193–8.
21. Kisler LB, Weissman-Fogel I, Coghill RC, Sprecher E, Yarnitsky D, Granovsky Y. Individualization of migraine prevention: a randomized controlled trial of psychophysical-based prediction of duloxetine efficacy. Clin J Pain 2019;35:753–65.
22. Fernandes C, Pidal-Miranda M, Samartin-Veiga N, Carrillo-de-la-Peña MT. Conditioned pain modulation as a biomarker of chronic pain: a systematic review of its concurrent validity. Pain 2019; 160:2679–90.
23. Hoegh M, Seminowicz DA, Graven-Nielsen T. Delayed effects of attention on pain sensitivity and conditioned pain modulation. Eur J Pain 2019;23:1850–62.
24. 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; 17:98–102.
25. Imai Y, Petersen KK, Mørch CD, Arendt Nielsen L. Comparing test-retest reliability and magnitude of conditioned pain modulation using different combinations of test and conditioning stimuli. Somatosens Mot Res 2016;33:169–77.
26. Cummins TM, McMahon SB, Bannister K. The impact of paradigm and stringent analysis parameters on measuring a net conditioned pain modulation effect: a test, retest, control study. Eur J Pain 2021;25:415–29.
27. Graven-Nielsen T, Izumi M, Petersen KK, Arendt-Nielsen L. User-independent assessment of conditioning pain modulation by cuff pressure algometry. Eur J Pain 2017;21:552–61.
28. Martel MO, Wasan AD, Edwards RR. Sex differences in the stability of conditioned pain modulation (CPM) among patients with chronic pain. Pain Med 2013;14:1757–68.
29. Cathcart S, Winefield AH, Rolan P, Lushington K. Reliability of temporal summation and diffuse noxious inhibitory control. Pain Res Manag 2009;14:433–8.
30. Oono Y, Nie H, Matos RL, Wang K, Arendt-Nielsen L. The inter- and intra-individual variance in descending pain modulation evoked by different conditioning stimuli in healthy men. Scand J Pain 2011;2:162–9.
31. 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. Reg Anesth Pain Med 2012;37:530–6.
32. 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 Muscoskel Disord 2013;14:182.
33. Wilson H, Carvalho B, Granot M, Landau R. Temporal stability of conditioned pain modulation in healthy women over four menstrual cycles at the follicular and luteal phases. Pain 2013; 154:2633–8.
34. Biurrun Manresa JA, Fritsche R, Vuilleumier PH, Oehler C, Mørch CD, Arendt-Nielsen L, et al. Is the conditioned pain modulation paradigm reliable? A test-retest assessment using the nociceptive withdrawal reflex. PLoS One 2014;9:e100241.
35. 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;19:93–6.
36. Granovsky Y, Miller-Barmak A, Goldstein O, Sprecher E, Yarnitsky D. CPM test-retest reliability: "standard" vs. "single test-stimulus" protocols. Pain Med 2016;17:521–9.
37. Haidich AB. Meta-analysis in medical research. Hippokratia 2010;14(1 Suppl):29–37.
38. Portney LG, Watkins MP. Foundations of clinical research applications to practice, 2nd ed. Upper Saddle River, New Jersey: Prentice Hall Health; 2000:61–77 pp.
39. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.

40. Bruton A, Conway JH, Holgate ST. Reliability: what is it, and how is it measured? Physiotherapy 2000;86:94–9.
41. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021;372:n160.
42. Mokkink LB, Boers M, van der Vleuten CPM, Bouter LM, Alonso J, Patrick DL, et al. COSMIN risk of bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Med Res Methodol 2020;20:293.
43. Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: a systematic review. J Rehabil Med 2011;43:23–31.
44. Kenny DA, Deborah K, Cook William L. Dyadic data analysis. New York, NY: Guilford; 2006.
45. Nuwailati R, Curatolo M, LeResche L, Ramsay DS, Spiekerman C, Drangsholt M. Reliability of the conditioned pain modulation paradigm across three anatomical sites. Scand J Pain 2020;20: 283–96.
46. Vuilleumier PH, Biurrun Manresa JA, Ghamri Y, Mlekusch S, Siegenthaler A, Arendt-Nielsen L, et al. Reliability of quantitative sensory tests in a low back pain population. Reg Anesth Pain Med 2015;40:665–73.
47. Gehling J, Mainka T, Vollert J, Pogatzki-Zahn EM, Maier C, Enax-Krumova EK. Short-term test-retest-reliability of conditioned pain modulation using the cold-heat-pain method in healthy subjects and its correlation to parameters of standardized quantitative sensory testing. BMC Neurol 2016;16:125.
48. Bossmann T, Brauner T, Lowak H, Anton F, Forster C, Horstmann T. Reliability of conditioned pain modulation for the assessment of endogenous pain control pathways. Neurol Psychiatr Brain Res 2016;22:155–61.
49. Costa YM, Morita-Neto O, de Araújo-Júnior EN, Sampaio FA, Conti PC, Bonjardim LR. Test-retest reliability of quantitative sensory testing for mechanical somatosensory and pain modulation assessment of masticatory structures. J Oral Rehabil 2017;44: 197–204.
50. Marcuzzi A, Wrigley PJ, Dean CM, Adams R, Hush JM. The long-term reliability of static and dynamic quantitative sensory testing in healthy individuals. Pain 2017;158:1217–23.
51. Lie MU, Matre D, Hansson P, Stubhaug A, Zwart JA, Nilsen KB. A tonic heat test stimulus yields a larger and more reliable conditioned pain modulation effect compared to a phasic heat test stimulus. Pain Rep 2017;2:e626.
52. Lie MU, Petriu E, Matre D, Hansson P, Andersen OK, Zwart JA, et al. Psychophysical or spinal reflex measures when assessing conditioned pain modulation? Eur J Pain 2019;23:1879–89.
53. Kennedy DL, Kemp HI, Wu C, Ridout DA, Rice ASC. Determining real change in conditioned pain modulation: a repeated measures study in healthy volunteers. J Pain 2020;21:708–21.
54. Alsouhibani A, Vaegter HB, Hoeger Bement M. Systemic exercise-induced hypoalgesia following isometric exercise reduces conditioned pain modulation. Pain Med 2019;20:180–90.
55. Larsen JB, Madeleine P, Arendt-Nielsen L. Development of a new bed-side-test assessing conditioned pain modulation: a test-retest reliability study. Scand J Pain 2019;19:565–74.
56. Naugle KM, Ohlman T, Wind B, Miller L. Test-Retest instability of temporal summation and conditioned pain modulation measures in older adults. Pain Med 2020;21:2863–76.
57. Jensen KB, Loitoile R, Kosek E, Petzke F, Carville S, Fransson P, et al. Patients with fibromyalgia display less functional connectivity in the brain's pain inhibitory network. Mol Pain 2012;8:32.
58. Sprenger C, Bingel U, Buchel C. Treating pain with pain: supraspinal mechanisms of endogenous analgesia elicited by heterotopic noxious conditioning stimulation. Pain 2011;152: 428–39.
59. Wilder-Smith CH, Schindler D, Lovblad K, Redmond SM, Nirkko A. Brain functional magnetic resonance imaging of rectal pain and activation of endogenous inhibitory mechanisms in irritable bowel syndrome patient subgroups and healthy controls. Gut 2004;53:1595–601.
60. Mutlu EK, Ozdincler AR. Reliability and responsiveness of algometry for measuring pressure pain threshold in patients with knee osteoarthritis. J Phys Ther Sci 2015;27:1961–5.
61. Tousignant-Laflamme Y, Pagé S, Goffaux P, Marchand S. An experimental model to measure excitatory and inhibitory pain mechanisms in humans. Brain Res 2008;1230:73–9.
62. Rezaii T, Hirschberg AL, Carlström K, Ernberg M. The influence of menstrual phases on pain modulation in healthy women. J Pain 2012;13:646–55.
63. Tousignant-Laflamme Y, Marchand S. Excitatory and inhibitory pain mechanisms during the menstrual cycle in healthy women. Pain 2009;146:47–55.
64. Smith YR, Stohler CS, Nichols TE, Bueller JA, Koeppe RA, Zubieta JK. Pronociceptive and antinociceptive effects of estradiol through endogenous opioid neurotransmission in women. J Neurosci 2006;26:5777–85.
65. Zubieta JK, Heitzeg MM, Smith YR, Bueller JA, Xu K, Xu Y, et al. COMT val158met genotype affects mu-opioid neurotransmitter responses to a pain stressor. Science 2003;299:1240–3.
66. Bjørkedal E, Flaten MA. Expectations of increased and decreased pain explain the effect of conditioned pain modulation in females. J Pain Res 2012;5:289–300.
67. Edwards RR, Fillingim RB, Ness TJ. Age-related differences in endogenous pain modulation: a comparison of diffuse noxious inhibitory controls in healthy older and younger adults. Pain 2003;101:155–65.
68. Riley JL 3rd, King CD, Wong F, Fillingim RB, Mauderli AP. Lack of endogenous modulation and reduced decay of prolonged heat pain in older adults. Pain 2010;150:153–60.
69. Larivière M, Goffaux P, Marchand S, Julien N. Changes in pain perception and descending inhibitory controls start at middle age in healthy adults. Clin J Pain 2007;23:506–10.
70. Lautenbacher S, Kunz M, Strate P, Nielsen J, Arendt-Nielsen L. Age effects on pain thresholds, temporal summation and spatial summation of heat and pressure pain. Pain 2005;115: 410–8.
71. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22:276–82.
72. Hróbjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol 2014;43: 1272–83.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/sjpain-2021-0149).