Title page

Absolute and relative reliability of a comprehensive quantitative sensory testing protocol in women treated for breast cancer

Lore Dams, Msc[1,2,3], Vincent Haenen, Msc[1,2], Elien Van der Gucht, Msc[1,2,3], Nele Devoogdt, PhD[2,4], Ann Smeets, MD, PhD[5], Koen Bernar, Msc[6], Tessa De Vrieze, PhD[2], An De Groef, PhD*[1,2,3+], Mira Meeus, Phd[1,3,7+]

[1]University of Antwerp, Faculty of Medicine and Health Sciences, Department of Rehabilitation Sciences and Physiotherapy, MOVANT, Antwerp, Belgium

[2]KU Leuven - University of Leuven, Department of Rehabilitation Sciences, Leuven, Belgium

[3]Pain In Motion International research group, www.paininmotion.be

[4]Department of Vascular Surgery and Department of Physical Medicine and Rehabilitation, Center for Lymphedema, UZ Leuven - University Hospitals Leuven, Leuven, Belgium.

[5]Department of Surgical Oncology, University Hospitals Leuven, Leuven, Belgium

[6]The Leuven Centre for Algology and Pain Management, University Hospitals Leuven, Leuven, Belgium

[7]Ghent University, Faculty of Medicine and Health Sciences, Department of Rehabilitation Sciences and Physiotherapy, Ghent, Belgium

+shared last author

**Support statement: ADG is a post-doctoral research fellow of the FWO-Flanders.*

Corresponding author

Lore Dams

Department of Rehabilitation Sciences and Physiotherapy

University of Antwerp

Campus Drie Eiken – Universiteitsplein 1, R.315

2610 Wilrijk, Belgium

0032 16 376 680

lore.dams@uantwerpen.be

Funding

The study is funded by Research Foundations – Flanders (FWO) (T005117N)

Conflict of interest

The authors have no conflicts of interest to declare.

Running title

Reliability QST protocol in breast cancer.

Abstract

Objective: Quantitative sensory testing (QST) are non-invasive psychophysical assessment techniques to evaluate functioning of the somatosensory nervous system. Despite the importance of reliability for correct use of QST results in research and clinical practice, the relative and absolute intra-and inter-rater reliability of a comprehensive QST protocol to evaluate the functioning of both peripheral and central somatosensory nervous system in a breast cancer population, has not yet been investigated.

Setting: University Hospitals, Leuven, Belgium.

Subjects: Thirty women at least six months after unilateral breast cancer surgery

Methods: The protocol included nine static and dynamic QST methods (mechanical detection-pain thresholds, pressure pain thresholds, thermal detection-pain thresholds for heat and cold, temporal summation and conditioned pain modulation (CPM)) performed in the surgical area and more distant regions. Absolute and relative intra (60-minutes interval) and inter-rater (one-week interval) reliability was evaluated using intraclass correlation coefficients, standard error of measurement and Bland-Altman plots.

Results: A moderate to excellent relative intra- and inter-rater reliability was found for the evaluation of mechanical thresholds, pressure pain thresholds and temporal summation. Reliability of the CPM paradigm was considered weak. Systematic bias between raters was noticed for detection of mechanical and cold stimuli at the non-affected trunk and CPM.

Conclusions: Except for the evaluation of CPM, the QST protocol was found suitable for identifying differences between subjects (relative reliability) and individual follow-up after breast cancer surgery (limited systematic bias) during a one-week timeframe.

Additional research is required to determine measurement properties that influence CPM test stability in order to establish a more reliable CPM test paradigm.

Key words: Breast Cancer, Quantitative Sensory Testing, Reliability

Introduction

In a significant proportion of women treated for breast cancer, somatosensory functioning is disturbed with high prevalence of sensory loss and/or pain (1). Quantitative sensory testing (QST) can be used to evaluate this (dys)function of the somatosensory nervous system (2) by administering standardized stimuli and quantifying the self-reported sensory experience (3, 4). Hence QST has already been applied to investigate whether somatosensory functioning before or immediately after breast cancer surgery is a predictor for pain in the long-term (5-7), to study the effect of a particular physical therapy intervention on somatosensory functioning (8, 9), and to inventory somatosensory profiles of women treated for breast cancer (10).

However, information regarding the reliability of QST is a prerequisite for its use in clinical-decision making. Firstly, reliability of a test instrument reflects the extent to which an observed test score is free from measurement error (i.e. *relative reliability* or degree to which individuals maintain their position in a sample over repeated measurements by the same (intra) or a different (inter) rater) (11). Secondly, reliability gives information about consistency of test scores across time, patients or observers assuming a stable response variable (i.e. *absolute reliability* or degree to which repeated measurements (inter or intra-rater) vary for individuals) (12, 13). Evaluating the reliability of a test instrument in a specific population is therefore essential and must be established before it can be used as a standard in both research and clinical practice (14).

Up to now, two studies investigated the absolute and relative intra-rater reliability of QST in a breast cancer population (15, 16). The QST protocol of Andersen et al.

demonstrated good relative intra-rater reliability for comparison of sensory function between participants, but less so for individual follow-up after breast cancer surgery. Evaluation of mechanical and thermal detection and pain thresholds both at affected and non-affected side was established, but evaluation of pressure pain thresholds was not incorporated (15). The study of Rasmussen et al. did examine the absolute and relative intra-rater reliability of pressure pain thresholds at the affected side and found a high relative intra-rater reliability. Evaluation of mechanical or thermal thresholds or pressure pain thresholds at the non-affected side were not incorporated in the protocol.

None of the two studies evaluated the reliability of mechanical or pressure pain thresholds at areas more distant from the breast surgery (15, 16). However, assessing mechanical thresholds at a remote body region may provide more information on the extra segmental spreading of pain sensitivity (17) and is therefore commonly used to evaluate central nociceptive processing in breast cancer research (8, 9, 18-20).

Additionally, none of the two studies on the reliability of QST in breast cancer incorporated the evaluation of two other methods to evaluate central nociceptive processing, namely conditioned pain modulation or temporal summation. Although, growing evidence supports the presence of an altered central nociceptive processing in a breast cancer population (6, 21-25) and the need to evaluate this phenomenon through dynamic QST methods of conditioned pain modulation and temporal summation (evaluating the response to a number of stimuli instead of one static sensory threshold) (26). Reliability studies for these QST methods are limited to non-cancer populations and results are inconclusive (27-29).

Therefore, the aim of the current study was to examine relative and absolute inter-and intra-rater reliability of a comprehensive QST protocol for evaluating peripheral and central somatosensory nervous system processing during a one-week timeframe in women at least 6 months after breast cancer surgery. Besides thermal and mechanical thresholds, the protocol also incorporated pressure pain thresholds as well as evaluation of central nociceptive processing by means of mechanical and thermal thresholds at areas more distant from the breast surgery, temporal summation and conditioned pain modulation.

Methods

Participants

Participants were included between March 2019 and November 2020 at the Department of Physical Medicine and Rehabilitation of the University Hospitals of Leuven campus Gasthuisberg (Belgium). All participants were randomly recruited from a cohort of women participating in a randomized controlled trial investigating the effectiveness of pain neuroscience education (EduCan trial, NCT03351075) (30). As a consequence, randomized controlled trial’s eligibility criteria also applied to the present study, requiring that the participants 1) were diagnosed with histologically confirmed invasive or non-invasive primary breast cancer, 2) had undergone one of the following surgeries: mastectomy including either a sentinel node biopsy or axillary lymph node dissection (with or without breast reconstruction) or breast conserving surgery including axillary lymph node dissection, 3) were female, 4) were 18 years or older, 5) comprehended the Dutch language (reading, listening, writing and speaking). Patients with bilateral breast cancer or active metastases were excluded. At time of inclusion in the present study, patients had undergone surgery six months prior. The study protocol was approved by the Ethical Committee of the University Hospitals Leuven (s60702) and all participants gave written informed consent prior to their enrollment.

Study design

QST was performed three times on each participant. The duration of one QST assessment was between 60 and 90 minutes. The first (A1) and second assessment (A2) took place on the same day, with a 60-minute break in between. The third assessment (A3) took place one week later.

The first (A1) and third assessment (A3) were executed by the same rater (LD) (intra-rater reliability). The second assessment (A2) was performed by different raters (FP or VH) (inter-rater reliability). All raters in this study performed QST in clinical routine. Before the start of the study, the QST-protocol was reiterated by the raters during two sessions. All raters were blinded to each other results. Though A1 and A3 were carried out by the same rater, results of A1 were not processed until all assessments had been completed.

The QST was executed in a quiet room with an approximate temperature between 21°C and 23°C. For each QST method, standardized test instructions were given prior to testing. Regarding the evaluation of pain thresholds, it was also emphasized that painful is not the moment from when a stimulation becomes unbearable for the participant, but rather from when a stimulation is perceived as unpleasant by the participant.

Participants were placed in a sitting position, with the lower arms supported by a table.

QST was performed at predefined anatomical locations (see Figure 1 for a detailed description of all test locations). When the predefined test location was located on a scar, the test was performed more proximally or distally from the scar (direction depending on where the predefined test location could be approached as closely as possible). For each participant, all test locations were marked prior to testing by the same rater (LD) with a dermatographic pencil. At the day of the first (A1) and second assessment (A2), test locations were marked for both sessions before the first assessment (A1). At the third assessment one week later (A3), test locations were indicated in the same way by the same rater (LD) prior to testing.

Nine different QST methods were included in the QST protocol. The protocol was based on the QST protocol as recommended by the German Research Network on Neuropathic Pain (DFNS) to evaluate somatosensory functioning comprehensively (31-33). However, the DFNS protocol did not include the evaluation of nociceptive processing by means of conditioned pain modulation. For this QST method, the current study used the protocol described by Granovsky et al. (2016) (14). See Table 1 for a comprehensive overview of all QST methods of the QST protocol and corresponding test locations. Four different sequences of QST methods were used to ensure that the results were not influenced by the measurement sequence. However, within one participant, the same sequence of examinations was used.

(*Suggestion to insert Figure 1 here*)

(*Suggestion to insert Table 1 here*)

Mechanical detection and mechanical pain threshold

The **mechanical detection threshold (MDT)** and **mechanical pain threshold (MPT)** were assessed using a standardized set of 12 nylon monofilaments (Optihair2-Set, Marstock Nervtest, Germany) exerting forces between 0.25 and 512 millinewton (mN). To determine thresholds, monofilaments were applied with a rate of 2 seconds on and 2 seconds off, in an ascending and descending sequence, starting with a force of 8mN.

The **MDT** was defined as the lowest mechanical force the participant could identify. Participants were instructed to keep their eyes closed and to verbally indicate when the touch was detected. Two consecutive forces had to be detected by the participant to rule out coincidence. The geometric mean of the result of the ascending (the first stimulus detected) and descending sequence (the last stimulus detected) was calculated (mN) (31-33).

The **MPT** was defined as the lowest mechanical force the participant perceived as painful. Participants were instructed to keep their eyes closed and to verbally indicate when the touch was experienced as painful. Two consecutive forces had to be detected by the participant to rule out coincidence. The geometric mean of the result of the ascending (the first stimulus perceived as painful) and descending sequence (the last stimulus experienced as painful) was calculated (mN) (31-33).

Pressure pain threshold

The pressure pain threshold (PPT) was measured by using a digital pressure algometer (Wagner FDX, Greenwich CT, USA) with a flat round rubber tip, probe area 1 cm².The PPT was defined as the amount of pressure by which the perception of pressure was first perceived as painful (31-33). It was determined by two series of ascending pressure at a rate of approximately 0.1 kgf/s until the participant indicated that the pressure was first perceived as painful by saying ‘stop’. The final threshold was the arithmetic mean of two trials (kgf) (21).

Thermal detection and thermal pain threshold

Thermal thresholds were determined using a computer-controlled thermode system (TSA II Medoc, Israel) with a 3 × 3 cm (9 cm²) thermode of Peltier elements. A method of limits protocol was applied by instructing the participant to push the computer-controlled button when they experienced a change from a thermo-neutral state to a distinct warm (**warmth detection threshold, WDT**) or cold sensation (**cold detection threshold, CDT**) (31-33).

In addition to the thermal detection thresholds, thermal pain thresholds were determined by instructing the participant to push the computer-controlled button when the sensation of warmth (**heat pain threshold, HPT**) or cold (**cold pain threshold, CPT**) first changes to being painful (31-33).

For each trial, baseline temperature of the probe was 32°C and temperature decreased/increased at a rate of 1°C/s. For safety reasons temperature increase and decrease was limited to 50°C and 0°C respectively. The final thermal detection and pain thresholds were defined as the arithmetic mean threshold temperature of three consecutive measurements (31-33).

Conditioned pain modulation

Conditioned pain modulation (CPM) was assessed using a computer-controlled thermode system (Q-sense Medoc, Israel) with two 3 × 3 cm (9 cm²) thermodes. A parallel CPM paradigm was applied in which an identical test stimulus was given before and then in parallel to the conditioning stimulus (14).

First, the intensity of the test stimulus was determined individually, based on the temperature required to evoke a painful sensation with a rating of four on a 0-10 numerical rating scale (NRS) (Pain4). Pain4 temperature was defined before the actual CPM paradigm by applying a series of heat stimuli at the volar side of the non-affected lower arm (location 10) and asking the participant to verbally rate the intensity of the perceived sensation following each stimulus on a 0-10 NRS with 0 representing no painful sensation and 10 the most painful sensation ever experienced. The baseline temperature was 32°C with an increasing rate of 2°C/s and decreasing rate of 1°C/s. During the first stimulation, baseline temperature increased up to 43°C. If an NRS score above or below 4/10 was given for a certain temperature, temperature of the next stimulation was decreased or increased with 1°C respectively. During the search for Pain4 temperature, a maximum of five heat stimulations was given. Minimum temperature of the test stimulus was 39°C, maximum temperature was 46°C.

After determination of the intensity of the test stimulus, the parallel CPM paradigm was started. During 45 seconds the individually determined noxious contact heat stimulus (temperature Pain4) was applied to the volar side of the affected lower arm (location 11). Participants were asked to rate verbally the intensity of the test stimulus during the course of the stimulus at 10 seconds, 20 seconds, 30 seconds and 40 seconds using a 0-10 NRS with 0 representing no painful sensation and 10 the most painful sensation ever experienced. After a 120 seconds break, the conditioning stimulus was applied at the volar side of the non-affected lower arm (location 10) for 65 seconds. The intensity for the conditioning stimulus was set 0.5°C above the intensity of the test stimulus. Twenty seconds after initiation of the conditioning stimulus, the test stimulus was applied in parallel (location 11) and verbal ratings of intensity were obtained at 10 seconds, 20 seconds, 30 seconds and 40 seconds of stimulation (0-10 NRS). CPM was calculated as a difference in NRS score between the conditioned test stimulus and the test stimulus without conditioning for every 10-seconds-long epoch.

Temporal summation

Temporal summation was examined by applying a train of pinprick stimuli and evaluating the perceived painfulness. After a single stimulation, a train of pinprick stimuli were given during 30 seconds with a stimulation force of 256mN (Optihair2-Set, Marstock Nervtest, Germany) at a rate of 1 stimulation/s. Participants were asked to rate the perceived sensation on a 0-10 NRS with 0 representing no painful sensation and 10 the most painful sensation ever experienced for the single stimulus, immediately after the train of pinprick stimulations as well as 15 seconds after the final stimulation. The difference between the NRS score immediately after the train of pinprick stimuli and the NRS score after the first stimulation was seen as outcome measure for temporal summation (34) .

Statistics

Descriptive statistics for continuous values are presented as mean - standard deviation (SD) for normally distributed data and median and interquartile range (IQR) for not normally distributed data. Categorical variables are presented as number and proportion (%).

Data of all QST methods were analyzed for their distribution properties according to the method of Rolke et al. (33). Skewness, kurtosis, and Kolmogorov-Smirnov *d* statistics were calculated for raw data and log-transformed data (base 10). The geometric mean of skewness and kurtosis was determined and multiplied by the Kolmogorov-Smirnov d for each distribution as a measure of goodness of fit to the normal distribution. Log-transformation was considered to be superior, when the ratio for raw data to log-transformed data exceeded a factor of 3. To avoid a loss of zero values, a small constant (0.1) was added to QST results of CDT and CPT before log-transformation (35).

To evaluate relative reliability, intraclass correlation coefficients (ICC) for average measurements were calculated to evaluate both relative inter- and intra-rater reliability. The ICC model was chosen a priori with a two-way mixed effects model for the intra-rater reliability (model 3) and a two-way random effects model for the inter-rater reliability (model 2). An ICC below 0.50 represented weak reliability, between 0.50 and 0.75 moderate reliability, between 0.75 and 0.90 good reliability and above 0.90 excellent reliability (12).

While the ICC gives information about the proportion of the observed variance that is a consequence of true variance between measurements (relative reliability), it gives no information about the response stability or magnitude of the disagreement between measurements (absolute reliability). Therefore, the standard error of measurement (SEM) was calculated in order to interpret the magnitude of the within subject variation.

Bland-Altman plots were constructed to visualize potential bias of the data and limits of agreement (LoA). The plots display differences between QST measurements (vertical axis) against the mean value of both measurements (horizontal axis). If differences are systematically greater or less than zero, the measurements are systematically greater or lower for a certain time or rater. In case of good agreement, the majority of the points lie within the 95% limits of agreement, with an even distribution of points on both sides of the mean difference to indicate no systematic bias (36). Evaluation of systematic bias was based on linear regression analysis and visual check of the Bland-Altman plot. Statistical analyses were performed using IBM SPSS Statistics for Windows version 27.0. The 0.01 level of significance was applied.

Results

Participants

A total of 30 women treated for breast cancer with a mean age of 57 were included. One participant had to cancel her appointments for the third evaluation (A3) because of illness. Consequently, the intra-rater reliability was evaluated for 29 participants. The vast majority of women completed primary treatment for breast cancer, except 4 participants who were still receiving radiotherapy. Patient characteristics are summarized in Table 2.

(*Suggestion to insert Table 2 here*)

Absolute and relative reliability

Mechanical detection and mechanical pain thresholds

Relative and absolute reliability of mechanical detection (MDT) and pain (MPT) thresholds is reported in Table 3.

ICC for MDT generally showed moderate to good relative intra- and inter-rater reliability (ICC range 0.556-0.792) at all test locations. Except for the trunk and quadriceps at the non-affected side where the relative inter-rater reliability was rather weak (ICC trunk 0.160, ICC Quadriceps 0.462).

ICC for MPT generally showed moderate relative intra- and inter-rater reliability (ICC range 0.504-0.740) at all test locations. Except for the upper arm and trunk at the non-affected side where the relative intra-rater reliability was rather weak (ICC upper arm

0.177, ICC trunk 0.374).

The analysis of the Bland-Altman plots showed a systematic bias for the evaluation of MDT between raters at the non-affected trunk. The systematic difference in MDT between rater 1 and rater 2 became greater for higher MDT (Supplementary material S1, Scatter plot MDT Trunk non-affected side A2-A1). No systematic bias was found for the MPT. See Supplementary material S1-S2 for individual Bland-Altman plots.

Pressure pain threshold

Relative and absolute reliability of pressure pain thresholds (PPT) is reported in Table 4. ICC for PPT generally showed good to excellent relative intra- and inter-rater reliability for all test locations (ICC range 0.762-0.916) except for the affected trunk and non-affected pectoralis region, which showed moderate relative inter-rater reliability (ICC trunk 0.631, ICC pectoralis 0.641). SEM values ranged from 0.217 to 1.110 kgf.

The analysis of the Bland-Altman plots showed a systematic bias for the evaluation of PPT by the same rater at the upper trapezius muscle at the affected side (Supplementary material S3, Scatter plot PPT Upper Trapezius affected side A3-A1). See Supplementary material S3 for individual Bland-Altman plots.

Thermal detection and thermal pain threshold

Relative and absolute reliability of thermal detection and pain thresholds is reported in Table 5. For test locations at the affected side (upper arm and trunk), ICC generally showed moderate to good relative intra- and inter-rater reliability (upper arm ICC range 0.606-0.872, trunk ICC range 0.685-0.791). An excellent relative intra-rater reliability was found for the evaluation of WDT at the affected arm (ICC 0.907) as well as for the evaluation of CPT at the affected trunk (ICC 0.917). For test locations at the non-affected side (upper arm, trunk and quadriceps), ICC generally showed weak to moderate relative intra- and inter-rater reliability for the evaluation of thermal detection thresholds (WDT and CDT) and moderate to good relative intra- and inter-rater reliability for the evaluation of thermal pain thresholds (HPT and CPT).

The analysis of the Bland-Altman plots showed a systematic bias for the evaluation of CDT between raters at the non-affected upper arm and trunk (Supplementary material S5, Scatter plot CDT Upper arm non-affected side A2-A1 and Scatter plot CDT Trunk non-affected side A2-A1). An increasing disagreement for the lower thresholds was noticed for both locations at the non-affected side. For the evaluation of thermal pain thresholds, no systematic bias was determined. See Supplementary material S4-S7 for individual Bland-Altman plots.

Conditioned pain modulation

Relative and absolute reliability of thermal detection and pain thresholds is reported in Table 6. A good relative inter- and intra-rater reliability was found for the evaluation of the temperature of test stimulus (ICC intra 0.877, ICC inter 0.859). For the evaluation of conditioned pain modulation itself, a weak to moderate relative intra- and inter-rater reliability was found (ICC range inter-rater reliability 0.052-0.695, ICC range intra-rater reliability 0.069-0.687).

The analysis of the Bland-Altman plots showed a systematic bias for the components evaluating the CPM effect (CPM 30s CPM 40s and CPM mean). The CPM effect systematically decreased between assessments by different raters on the same day (Supplementary material S8, Scatter plot CPM 30s A2-A1, Scatter plot CPM 40s A2-A1, Scatter plot CPM mean A2-A1). See Supplementary material S8 for individual Bland-Altman plots.

Temporal summation

Relative and absolute reliability of temporal summation is reported in Table 7. ICC generally showed moderate to good relative intra- and inter-rater reliability for all components of the evaluation of temporal summation.

Based on the analysis of the Bland-Altman plots no systematic bias was determined. See Supplementary material S9 for individual Bland-Altman plots.

(*Suggestion to insert Table 3-7 here*)

Discussion

This is the first study evaluating both relative and absolute inter- and intra-rater reliability of a comprehensive QST protocol consisting of nine different QST methods in a population of women treated for breast cancer.

For the evaluation of **mechanical detection (MDT) and mechanical pain (MPT) thresholds**, the same moderate relative intra-rater reliability was found as in the study of Andersen et al., that evaluated absolute and relative intra-rater reliability of these QST methods in women one year after surgery for breast cancer (15). Only for the evaluation of MPT at the unaffected side, a lower relative reliability was noticed in the current study. This was likely due to variation in test location between studies (upper lateral quadrant of the breast versus midaxillary line).

Up to now, the inter-rater reliability of the protocol for MDT and MPT used in this study has not been investigated in a population of breast cancer and is hardly studied in general. Moderate inter-rater reliability was found in the current study for MDT and MPT, although the usage of handheld filaments has been criticized, primarily due to the possible variability of the application procedure (degree of filament indention or unintentional movement of the hand) and concerns that the characteristics of the mechanical filament may alter with time (37-39). Based on the Bland-Altman plots for MDT and MPT, these methods were found to be less suitable for individual follow-up in the study of Andersen et al. (15). However, findings of the current study indicate that MDT and MPT are reliable methods for individual follow-up after breast cancer surgery.

Regarding the evaluation of **pressure pain thresholds (PPT)** good to excellent relative intra-rater reliability was found for both affected as non-affected side. Compared to the study of Rasmussen et al., more variability in absolute and relative intra-rater reliability was found for PPT at the upper trapezius muscle at the affected side (16). A possible explanation can be the difference between study populations. In the study of Rasmussen et al., participants had to report pain of at least 3/10 in the upper limb, while in the current study none of the participants had pain in the upper limb at time of assessment.

Both for the evaluation of PPT and thermal thresholds, the relative inter-rater reliability was always lower than the intra-rater reliability. Some studies suggest that the number of tests can contribute to habituation or sensitization to testing, as well as to concentration during testing, thereby having the ability to increase or decrease thresholds between sessions (40-42). Despite the fact that participants had been through the whole QST test protocol for at least three times, the method of limits was used for evaluation of both thermal and pressure pain thresholds with important implications of participant’s reaction time (43, 44). Although the raters were trained to be competent in QST, the second rater had less experience than the first rater. However, the influence of extent of examiner experience (above competency) on reliability of QST is currently not known.

Another finding concerning the reliability of **thermal thresholds** was the lower relative intra-rater reliability at non-affected side compared to affected side for the detection of warm and cold stimuli as well as for the evaluation of painful heat stimuli. This was also noticed a study of Geber et al. where in patients with unilateral neuropathic disorder, repeatability of thermal and mechanical testing for the affected area seemed to be better than for the unaffected side (39). A suggested hypothesis is a lack of systematic somatosensory variance in unaffected areas (39). In addition, attentional changes to uncomfortable or deafferented regions of the body can also contribute to a better reliability in clinically affected areas (45).

Beside side to side differences in intra-rater reliability for thermal thresholds, studies in other populations describe a lower intra-rater reliability of thermal pain thresholds measurements, particularly for cold pain (29). However, this finding could not be confirmed in the current study, as the reliability of cold pain thresholds generally ranged from good to perfect and was always higher than the reliability of the heat pain thresholds. A possible explanation for this discrepancy can be the handling of data from participants that did not report a heat or cold pain sensation within the applied temperature limits. When the thresholds could not be detected within the temperature limits, the maximum value was recorded as test score in the current study. While this approach was also applied in the studies of Andersen et al.(15) and Felix et al.(46), other studies on reliability of thermal thresholds describe exclusion of this data from analysis in case a temperature limit was reached and the subject did not report any sensation (38, 47, 48). The latter method may decrease number and range of test scores with possible implications on ICC magnitude (49).

While growing evidence supports the presence of altered central nociceptive processing in a population of breast cancer patients and survivors (6, 21-25), this study is the first one to examine the reliability of dynamic QST methods evaluating this phenomenon in a cancer population. Regarding **temporal summation (TS)** both inter- and intra-rater reliability were found moderate to good. This in contrast to the evaluation of **conditioned pain modulation (CPM)**, for which relative inter- and intra-rater reliability were determined weak to moderate. These results are in line with the study of Granovsky et al. that investigated the same two-thermode parallel heat design as in the current study and evaluated its relative intra-rater reliability (one-week time interval) in healthy participants (14). Although reliability ranged from weak to moderate for the evaluation of test and conditioned test stimuli, reliability for the components evaluating the CPM effect (CPM 10-40s and CPM mean) was remarkably weak. The following explanations can be given for this result. First, previous studies suggested that the stability of the CPM-effect may be enhanced by a greater stimulus intensity because in this case habituation would be greater (50, 51). In the study of Granovsky et al. (14) a temperature that evoked pain of at least 30/100 was defined as temperature for the test stimulus, in the current study a rating of 4/10 was used. Second, Granovsky et al. suggested that a shorter duration of the CPM procedure may have better impact on CPM reliability (14). Third, Kennedy et al. stated that the repeatability of different test and conditioning stimuli differs across sessions, and this lack of repeatability of the components of the CPM paradigm may decrease the repeatability for the sum of the paradigm (27). In addition, CPM was calculated as a difference between test scores, restricting inter-individual range of values with a possible impact on ICC magnitude (27, 52).

Regarding inter-rater reliability Bland-Altman plots showed a systematic bias for inter-rater agreement for evaluation of the CPM effect. More specifically CPM effect systematically decreased between assessments on the same day. The systematic decrease in CPM effect may be explained by a particular form of learning effect, whereby experience with the test and conditioning stimulus from the previous session that day could lead to a decreased threat value for these stimuli, reducing the magnitude of the decreasing inhibitory reaction (52).

Strengths of the current study were first the evaluation of both relative as well as absolute reliability of a comprehensive QST protocol, with QST methods mainly according to the recommendations of the Deutscher Forschungsverbund Neuropathischer Schmerz (DFNS) (33) and including evaluation of CPM. Furthermore, measurement error (and according reliability of an instrument) was controlled as much as possible by training of the raters and standardization of the protocol, by using the same test instruments and room for both inter- as intra rater reliability testing and by using the mean of two or more trials to reduce overall error in test scores. Finally, QST will quantify somatosensory function based on subjective (psychophysical) methods where participants need to report their sensory experiences. Consequently, consistency in QST results is also dependent on cooperation of the participant, perception, physical or emotional status etc. (53). Somatosensory functioning and in particular pain, is a complicated and challenging experience to quantify and it is extremely challenging to achieve equivalent measures of pain thresholds for two separate test occasions (54). It is likely that there will be some difference in the results over time due to the subjective aspect of the experience being evaluated (38). To counter this to a certain extent, the current study used a two-way mixed effects model (ICC 3,1) to evaluate the intra-rater reliability. In this model a systematic difference is allowed between measurements.

Nevertheless, some **limitations** of the current study need to be acknowledged. First, a time interval of one week was chosen for evaluation of intra-rater reliability based on previous studies examining intra-rater reliability of QST in a population of breast cancer (15, 16). Although results of other studies in non-cancer populations demonstrate more consistency for shorter time intervals (duration in days rather than longer time intervals) (39, 47). For the evaluation of inter-rater reliability, a time interval of one hour was chosen for practical reasons reducing the time the participant needed to stay in the hospital. It is not certain that this time interval was sufficient to allow a washout period of previous test results. Second, a formal power calculation was also not carried out. Third, all study participants completed primary treatment for breast cancer except four participants who were still receiving radiotherapy. This may have affected stability of somatosensory behavior while reliability is actually the extent to which a value can be obtained during repeated assessment of unchanging behavior. However, variation in these participants was similar to other participants.

Conclusion

The present study evaluated both absolute and relative reliability of a comprehensive QST protocol for evaluating peripheral and central somatosensory nervous system processing over a one-week timeframe in women at least 6 months after breast cancer surgery. With exception of CPM, the QST protocol was found to be suitable in this population for identifying differences between subjects (relative reliability) as well as for individual follow-up after breast cancer surgery (limited systematic bias) during the aforementioned timeframe. Overall, it the evaluation of pressure pain thresholds and temporal summation appeared to be the most consistent. Additional research is required to determine measurement properties that influence CPM test stability in order to establish a more reliable CPM test paradigm.

Acknowledgements

The authors thank Frauke Penen for assessment of participants as well as Geert Verbeke and Annouschka Laenen for the statistical advice. Finally, they thank all participants who took part in this study.

Conflict of Interest – Disclosure summary

The study is funded by Research Foundations – Flanders (FWO) (T005117N). All authors critically revised the manuscript for important intellectual content and approved the final manuscript. The authors have no conflicts of interest to declare.

References

1. Mejdahl MK, Andersen KG, Gartner R, Kroman N, Kehlet H. Persistent pain and sensory disturbances after treatment for breast cancer: six year nationwide follow-up study. Bmj. 2013;346:f1865

2. Backonja MM, Attal N, Baron R et al. Value of quantitative sensory testing in neurological and pain disorders: NeuPSIG consensus. Pain. 2013;154(9):1807-19.

3. Verberne WR, Snijders TJ, Liem KS et al. [Applications of 'quantitative sensory testing']. Nederlands Tijdschrift voor Geneeskunde. 2013;157(5):A5434.

4. Hall T, Briffa K, Schafer A et al. Quantitative Sensory Testing: Implications for clinical practice. In: Hall T, Briffa K, Schäfer A et al., eds. Grieve's Modern Musculoskeletal Physiotherapy: Vertebral Column and Peripheral Joints 4. UK: Elsevier Health Sciences; 2015.

5. Andersen KG, Duriaud HM, Kehlet H, Aasvang EK. The Relationship Between Sensory Loss and Persistent Pain 1 Year After Breast Cancer Surgery. J of Pain. 2017;18(9):1129-38.

6. Schreiber KL, Martel MO, Shnol H et al. Persistent pain in postmastectomy patients: comparison of psychophysical, medical, surgical, and psychosocial characteristics between patients with and without pain. Pain. 2013;154(5):660-8.

7. La Cesa S, Sammartino P, Mollica C et al. A longitudinal study of painless and painful intercostobrachial neuropathy after breast cancer surgery. Neurol Sci. 2018;39(7):1245-51.

8. Cantarero-Villanueva I, Fernandez-Lao C, Fernez-de-Las-Penas C et al. Effectiveness of water physical therapy on pain, pressure pain sensitivity, and myofascial trigger points in breast cancer survivors: a randomized, controlled clinical trial. Pain Med. 2012;13(11):1509-19.

9. Fernandez-Lao C, Cantarero-Villanueva I, Fernez-de-Las-Penas C et al. Effectiveness of a multidimensional physical therapy program on pain, pressure hypersensitivity, and trigger points in breast cancer survivors: a randomized controlled clinical trial. Clin J Pain. 2012;113-21 p.

10. Mustonen L, Vollert J, Rice ASC, Kalso E, Harno H. Sensory profiles in women with neuropathic pain after breast cancer surgery. Breast Cancer Res Treat. 2020;182(2):305-15.

11. Baumgartner TA. Norm-referenced measurement: reliabilty. In: Safrit MJW, Wood TM,

eds. Measurement Concepts in Physical Education and Exercise Science (pp.45-72). Champaign, Illinois,1989.

12. Portney LG. Foundations of Clinical Research: Applications to Evidence-Based Practice, 4th ed. Philadelphia: F.A. Davis Company, 2020.

13. Moloney NA, Hall TM, Doody CM. Reliability of thermal quantitative sensory testing: a systematic review. J Rehabil Res Dev. 2012;49(2):191-207.

14. Granovsky Y, Miller-Barmack A, Goldstein O, Sprecher E, Yarnitsky D. CPM Test-Retest Reliability: "Standard" vs "Single Test-Stimulus" Protocols. Pain Med. 2016;17(3):521-9.

15. Andersen KG, Kehlet H, Aasvang EK. Test-retest agreement and reliability of quantitative sensory testing 1 year after breast cancer surgery. Clin J Pain. 2015;31(5):393-403.

16. Rasmussen GHF, Kristiansen M, Arroyo-Morales M, Voigt M, Madeleine P. Absolute and relative reliability of pain sensitivity and functional outcomes of the affected shoulder among women with pain after breast cancer treatment. PLoS One. 2020;15(6):e0234118.

17. Arendt-Nielsen L, Morlion B, Perrot S et al. Assessment and manifestation of central sensitisation across different chronic pain conditions. Eur J Pain. 2018;22(2):216-41.

18. Caro-Moran E, Diaz-Rodriguez L, Cantarero-Villanueva I et al. Nerve pressure pain hypersensitivity and upper limb mechanosensitivity in breast cancer survivors: a case-control study. Pain Med. 2014;15(10):1715-23.

19. Fernandez-Lao C, Cantarero-Villanueva I, Fernandez-de-las-Penas C et al. Myofascial Trigger Points in Neck and Shoulder Muscles and Widespread Pressure Pain Hypersensitivtiy in Patients With Postmastectomy Pain: Evidence of Peripheral and Central Sensitization. Clin J Pain. 2010;26(9):798-806.

20. Fernandez-Lao C, Cantarero-Villanueva I, Fernandez-de-las-Penas C et al. Widespread mechanical pain hypersensitivity as a sign of central sensitization after breast cancer surgery: comparison between mastectomy and lumpectomy. Pain Med. 2011;12:72–78.

21. Edwards RR, Mensing G, Cahalan et al. Alteration in pain modulation in women with persistent pain after lumpectomy: influence of catastrophizing. J Pain Symptom Manage. 2013;46(1):30-42.

22. Henry NL, Conlon A, Kidwell KM et al. Effect of estrogen depletion on pain sensitivity in aromatase inhibitor-treated women with early-stage breast cancer. J Pain. 2015;15(5):468-75.

23. Palmer ACS, Souza A, Dos Santos VS et al. The Effects of Melatonin on the Descending Pain Inhibitory System and Neural Plasticity Markers in Breast Cancer Patients Receiving Chemotherapy: Randomized, Double-Blinded, Placebo-Controlled Trial. Front Pharmacol. 2019;10:1382.

24. Gottrup HA, Andersen J, Arendt-Nielsen L, Jensen TS. Psychophysical examination in patients with post-mastectomy pain. Pain. 2000;87(3):275-84.

25. Vilholm OJ, Cold S, Rasmussen L, Sindrup SH. Sensory function and pain in a population of patients treated for breast cancer. Acta Anaesthesiol Scand. 2009;53(6):800-6.

26. Arendt-Nielsen L, Yarnitsky D. Experimental and clinical applications of quantitative sensory testing applied to skin, muscles and viscera. J Pain. 2009;10(6):556-72.

27. Kennedy DL, Kemp HI, Ridout D, Yarnitsky D, Rice AS. Reliability of conditioned pain modulation: a systematic review. Pain. 2016;157(11):2410-9.

28. 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(11):2863-2876.

29. Middlebrook N, Heneghan NR, Evans DW, Rushton A, Falla D. Reliability of temporal summation, thermal and pressure pain thresholds in a healthy cohort and musculoskeletal trauma population. PLoS One. 2020;15(5):e0233521.

30. De Groef A, Devoogdt N, Van der Gucht E et al. EduCan trial: study protocol for a randomised controlled trial on the effectiveness of pain neuroscience education after breast cancer surgery on pain, physical, emotional and work-related functioning. BMJ Open. 2019;9(1):e025742.

31. Rolke R, Baron R, Maier C et al. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): standardized protocol and reference values. Pain. 2006;123(3):231-43.

32. Mucke M, Cuhls H, Radbruch L et al. Quantitative sensory testing (QST). English version. Schmerz. 2016.

33. Rolke R, Magerl W, Campbell KA et al. Quantitative sensory testing: a comprehensive protocol for clinical trials. Eur J Pain. 2006;10(1):77-88.

34. Cathcart S, Winefield AH, Rolan P, Lushington K. Reliability of temporal summation and diffuse noxious inhibitory control. Pain Res Manag. 2009;14(6):433-8.

35. Magerl W, Wilk SH, Treede RD. Secondary hyperalgesia and perceptual wind-up following intradermal injection of capsaicin in humans. Pain. 1998;74(2-3):257-68.

36. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-10.

37. Arezzo J, Bolton C, Boulton A et al.Quantitative sensory testing: a consensus report from the Peripheral Neuropathy Association. Neurology. 1993;43(5):1050-2.

38. Wylde V, Palmer S, Learmonth ID, Dieppe, P. Test-retest reliability of Quantitative Sensory Testing in knee osteoarthritis and healthy participants. Osteoarthritis Cartilage. 2011;19(6):655-8.

39. Geber C, Klein T, Azad S et al. Test-retest and interobserver reliability of quantitative sensory testing according to the protocol of the German Research Network on Neuropathic Pain (DFNS): a multi-centre study. Pain. 2011;152(3):548-56.

40. May A, Rodriguez-Raecke R, Schulte A et al. Within-session sensitization and between-session habituation: a robust physiological response to repetitive painful heat stimulation. Eur J Pain. 2012;16(3):401-9.

41. Breimhorst M, Hondrich M, Rebhorn C, May A, Birklein F. Sensory and sympathetic correlates of heat pain sensitization and habituation in men and women. Eur J Pain. 2012;16(9):1281-92.

42. Jürgens TP, Sawatzki A, Henrich F, Magerl W, May A. An improved model of heat-induced hyperalgesia--repetitive phasic heat pain causing primary hyperalgesia to heat and secondary hyperalgesia to pinprick and light touch. PLoS One. 2014;9(6):e99507.

43. Yarnitsky D, Orchoa J L. Warm and cold specific somatosensory systems. Psychophysical thresholds, reaction times and peripheral conduction velocities. Brain. 1991;114 (Pt 4):1819-26.

44. Heldestad V, Linder J, Sellersjö L, Nordh E. Reproducibility and influence of test modality order on thermal perception and thermal pain thresholds in quantitative sensory testing. Clin Neurophysiol. 2010;121(11):1878-85.

45. Seminowicz DA, Davis KD. A re-examination of pain-cognition interactions: implications for neuroimaging. Pain. 2007;130(1-2):8-13.

46. Felix ER, Widerström-Noga EG. Reliability and validity of quantitative sensory testing in persons with spinal cord injury and neuropathic pain. J Rehabil Res Dev. 2009;46(1):69-83.

47. Yarnitsky D, Sprecher E, Zaslansky R, Hemli JA. Heat pain thresholds: normative data and repeatability. Pain. 1995;60(3):329-32.

48. Moloney NA, Hall TM, O'Sullivan TC, Doody CM. Reliability of thermal quantitative sensory testing of the hand in a cohort of young, healthy adults. Muscle Nerve. 2011;44(4):547-52.

49. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19(1):231-40.

50. Granot M, Weissman-Fogel I, Crispel Y et al. 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):142-9.

51. 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(12):2633-8.

52. 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(7):1217-23.

53. Chong PS, Cros DP. Technology literature review: quantitative sensory testing. Muscle Nerve. 2004;29(5):734-47

54. Gooberman-Hill R, Woolhead G, Mackichan F et al. Assessing chronic joint pain: lessons from a focus group study. Arthritis Rheum. 2007;57(4):666-71.

Legends of figures

Figure 1. Test locations quantitative sensory testing protocol

Legends of tables

Table 1. Overview quantitative sensory testing protocol

Table 2. Patient Characteristics

Table 3. Intraclass correlations, means, standard error of measurement and limits of agreement for mechanical detection and pain thresholds

Table 4. Intraclass correlations, means, standard error of measurement and limits of agreement for pressure pain thresholds

Table 5. Intraclass correlations, means, standard error of measurement and limits of agreement for thermal detection and pain thresholds

Table 6. Intraclass correlations, means, standard error of measurement and limits of agreement for the evaluation of conditioned pain modulation

Table 7. Intraclass correlations, means, standard error of measurement and limits of agreement for the evaluation of temporal summation

Table 1. *Overview quantitative sensory testing protocol*

QST Method	Device	Outcome	Test location (see Figure 1)
Mechanical detection threshold	Von Frey monofilaments	Geometric mean first and last detected stimulus (mN)	Inner upper arm (5-6) Lateral trunk (7-8) Quadriceps (9)
Mechanical pain threshold	Von Frey monofilaments	Geometric mean first and last painful stimulus (mN)	Inner upper arm (5-6) Lateral trunk (7-8) Quadriceps (9)
Pressure pain threshold	Digital algometer	Method of limits: arithmetic mean 2 trials (kgf)	Upper trapezius (1-2) Pectoral region (3-4) Lateral trunk (7-8) Quadriceps (9)
Thermal detection threshold -Warmth detection threshold -Cold detection threshold	Thermode system (TSA II)	Method of limits: arithmetic mean 3 trials (°C)	Inner upper arm (5-6) Lateral trunk (7-8) Quadriceps (9)
Thermal pain threshold -Heat pain threshold -Cold pain threshold	Thermode system (TSA II)	Method of limits: arithmetic mean 3 trials (°C)	Inner upper arm (5-6) Lateral trunk (7-8) Quadriceps (9)
Temporal summation	Von Frey monofilament 256 mN	Difference in pain intensity immediately after 30s stimulation and after single stimulation (NRS)	Pectoral region (3)
Conditioned pain modulation	Two-thermode system (Q-sense)	Difference in pain intensity conditioned test stimulus and test stimulus without conditioning (NRS)	Volar side lower arm -conditioning stimulus: non-affected side (10) - test stimulus: affected side (11)

kgf = kilogram-force, mN = millinewton, NRS = numerical rating scale, s = seconds

Table 2. *Patient characteristics - Numbers (%) are given unless specified otherwise (n = 30).*

Age (years) Mean (SD, range)	57 (10, 33-78)
BMI (kg/m²) Mean (SD)	26 (5)
Time between A1-A2 and A3 (days) Median (IQR)	6 (1)
Time since surgery (days) Median (IQR)	238 (176)
Type of breast surgery	
Mastectomy	28 (93%)
Breast conserving surgery	2 (7%)
Type of axillary surgery	
Sentinel lymph node biopsy	13 (43%)
Axillary lymph node dissection	17 (57%)
Surgery at dominant side	11 (37%)
Tumor size (histopathological staging)	
T0	8 (27%)
T1	9 (30%)
T2	7 (23%)
T3	4 (13%)
T4	2 (7%)
Lymph node stage (histopathological staging)	
N0	14 (47%)
N1	10 (33%)
N2	5 (17%)
N3	1 (3%)
Radiotherapy	25 (83%)
Breast region	2 (7%)
Thorax	22 (73%)
Median subclavian and parasternal nodes	25 (83%)
Axilla region	2 (7%)
Hormone Therapy (ongoing)	21 (70%)
Chemotherapy	24 (80%)
Target therapy (Herceptin) (ongoing)	7 (23%)
Pain	
Mean global pain intensity last week (VAS)	
VAS = 0	6 (20%)
VAS 1-10	3 (10%)
VAS 11-30	12 (40%)
VAS 31-40	4 (13%)
VAS >40	5 (17%)
Pain in area of surgery at assessment	0 (0%)

Abbreviations: A1 = first assessment (rater 1), A2 = second assessment (rater 2), A3= third assessment (rater 1), IQR = interquartile range, SD = standard deviation, VAS = visual analog scale

Table 3. *Intraclass correlations, means, standard error of measurement and limits of agreement for mechanical detection and pain thresholds*

		Intra-rater reliability						Inter-rater reliability						
Site	*Mod*	N	A1 Mean (SD)	A3 Mean (SD)	ICC (95% CI)	SEM	LoAs Lower to upper	N		A1 Mean (SD)	A2 Mean (SD)	ICC (95% CI)	SEM	LoAs Lower to upper
Upper arm A	MDT$_{log}$	29	0.389 (0.972)	0.451 (1.088)	0.716 * (0.478-0.856)	0.549	-1.46 to 1.59	30	MDT$_{log}$	0.356 (0.972)	0.702 (1.051)	0.666 (0.390-0.829)	0.584	-1.19 to 1.89
	MPT (mN)	29	371.670 (160.371)	357.988 (175.959)	0.504 (0.174-0.732)	118.434	-342.50 to 315.07	30	MPT (mN)	376.376 (159.650)	349.419 (164.358)	0.592 (0.303-0.782)	103.480	-313.86 to 259.95
Upper arm NA	MDT$_{log}$	29	-0.021 (0.604)	-0.135 (0.541)	0.792 (0.604-0.897)	0.261	-0.84 to 0.61	30	MDT$_{log}$	-0.040 (0.603)	0.070 (0.496)	0.644 (0.381-0.812)	0.328	-0.80 to 1.02
	MPT (mN)	29	317.885 (165.196)	302.150 (145.956)	0.177 (-0.196-0.506)	141.137	-407.64 to 376.17	30	MPT (mN)	324.355 (166.147)	301.085 (142.883)	0.619 (0.341-0.798)	95.375	-288.73 to 242.19
Trunk A	MDT$_{log}$	29	1.349 (1.145)	1.224 (1.316)	0.650 (0.377-0.819)	0.739	-2.15 to 1.90	30	MDT$_{log}$	1.329 (1.130)	1.597 (1.197)	0.673 (0.391-0.825)	0.665	-1.55 to 2.08
	MPT (mN)	29	448.470 (126.860)	404.600 (153.696)	0.704 (0.459-0.849)	76.319	-256.55 to 168.81	29	MPT (mN)	448.470 (126.860)	414.453 (158.552)	0.663 (0.402-0.825)	82.843	-261.26 to 193.22
Trunk NA	MDT$_{log}$	29	-0.317 (0.399)	-0.363 (0.407)	0.556 * (0.244-0.764)	0.268	-0.79 to 0.70	30	MDT$_{log}$	-0.326 (0.395)	0.055 (0.682)	0.160 * (-0.135-0.457)	0.493	-1.01 to 1.77
	MPT (mN)	29	354.169 (152.105)	351.681 (144.300)	0.374 (0.014-0.647)	117.258	-327.74 to 322.76	30	MPT (mN)	359.430 (152.212)	344.162 (166.156)	0.713 * (0.481-0.852)	85.279	-253.28 to 222.74
Qceps	MDT$_{log}$	29	0.182 (0.605)	0.228 (0.656)	0.603 (0.310-0.792)	0.397	-1.05 to 1.15	30	MDT$_{log}$	0.155 (0.612)	0.269 (0.713)	0.462 (0.130-0.701)	0.486	-1.24 to 1.46
	MPT (mN)	29	354.580 (147.062)	370.942 (145.034)	0.567 (0.260-0.771)	96.103	-249.89 to 282.61	30	MPT (mN)	359.827 (147.335)	367.357 (154.882)	0.740 (0.521-0.867)	77.050	-208.32 to 223.38

Abbreviations: A = affected side, A1 = first assessment (rater 1), A2 = second assessment (rater 2), A3= third assessment (rater 1), CI = Confidence Interval, ICC = intraclass correlation coefficient, LoAs = Limits of agreement according to Bland-Altman, MDT = mechanical detection threshold, mN = millinewton, MPT = mechanical pain threshold, N = number of participants, NA = non-affected side, SEM = standard error of measurement, Qceps = Quadriceps

* important influence of outlier on result (see Supplementary material S10 for results without outliers)

ICC > 0.90 = excellent reliability, 0.75-0.90 good reliability, 0.50-0.75 moderate reliability and < 0.50 weak reliability (12)

Table 4. *Intraclass correlations, means, standard error of measurement and limits of agreement for pressure pain thresholds (kgf)*

		Intra-rater reliability						**Inter-rater reliability**				
Site	N	A1 Mean (SD)	A3 Mean (SD)	ICC (95% CI)	SEM	LoAs Lower to upper	N	A1 Mean (SD)	A2 Mean (SD)	ICC (95% CI)	SEM	LoAs Lower to upper
Pect A	29	1.431 (0.635)	1.487 (0.617)	0.880 (0.761-0.942)	0.217	-0.54 to 0.66	30	1.462 (0.647)	1.585 (0.753)	0.762 (0.561-0.879)	0.341	-0.82 to 1.06
Pect NA	29	1.419 (0.522)	1.512 (0.530)	0.822 (0.655-0.912)	0.222	-0.52 to 0.71	30	1.452 (0.544)	1.696 (0.690)	0.641 (0.339-0.817)	0.370	-0.72 to 1.21
Trunk A	29	1.630 (1.526)	1.754 (1.509)	0.849 (0.704-0.926)	0.590	-1.51 to 1.76	30	1.733 (1.602)	1.867 (1.454)	0.631 (0.332-0.797)	0.928	-2.46 to 2.73
Trunk NA	29	1.383 (0.610)	1.447 (0.632)	0.821 (0.556-0.888)	0.263	-0.66 to 0.79	30	1.420 (0.630)	1.440 (0.759)	0.820 (0.656-0.910)	0.295	-0.81 to 0.85
UT A	29	2.287 (1.041)	2.521 (1.371)	0.834 (0.676-0.918)	0.491	-1.14 to 1.61	30	2.311 (1.031)	2.646 (1.285)	0.809 (0.590-0.910)	0.552	-0.97 to 1.64
UT NAS	29	2.535 (1.177)	2.720 (1.317)	0.916 (0.829-0.960)	0.361	-0.82 to 1.19	30	2.602 (1.215)	2.831 (1.406)	0.849 (0.704-0.925)	0.509	-1.14 to 1.60
Qceps	29	4.758 (2.436)	4.668 (2.367)	0.903 (0.805-0.953)	0.748	-2.16 to 1.98	30	4.833 (2.429)	5.271 (2.600)	0.805 (0.631-0.902)	1.110	-2.58 to 3.46

Abbreviations: A = affected side, A1 = first assessment (rater 1), A2 = second assessment (rater 2), A3 = third assessment (rater 1), CI = Confidence Interval, ICC = intraclass correlation coefficient, kgf = kilogram-force, LoAs = Limits of agreement according to Bland-Altman, N = number of participants, NA = non-affected side, SD = standard deviation, SEM = standard error of measurement, Qceps = Quadriceps.

ICC > 0.90 = excellent reliability, 0.75-0.90 good reliability, 0.50-0.75 moderate reliability and < 0.50 weak reliability (12)

Table 5. *Intraclass correlations, means, standard error of measurement and limits of agreement for thermal detection and pain thresholds*

		Intra-rater reliability						**Inter-rater reliability**					
Site	*Mod*	N	A1 Mean (SD)	A3 Mean (SD)	ICC (95% CI)	SEM	LoAs Lower to upper	N	A1 Mean (SD)	A2 Mean (SD)	ICC (95% CI)	SEM	LoAs Lower to upper
Upper arm A	WDT (°C)	29	39.428 (5.406)	39.038 (5.692)	0.907 (0.813-0.955)	1.692	-5.07 to 4.29	30	39.438 (5.312)	40.837 (6.093)	0.781 (0.579-0.891)	2.669	-5.70 to 8.50
	CDT (°C)	29	23.617 (10.384)	23.778 (10.055)	0.872 (0.746-0.938)	3.656	-9.97 to 10.30	30	23.629 (10.204)	23.457 (10.750)	0.661 (0.396-0.823)	6.100	-17.27 to 16.93
	HPT (°C)	29	45.408 (3.503)	45.651 (3.448)	0.606 (0.313-0.793)	2.181	-5.80 to 6.29	30	45.512 (3.489)	46.994 (3.045)	0.657 * (0.312-0.835)	1.913	-3.33 to 6.30
	CPT (°C)	29	13.175 (10.288)	11.842 (10.314)	0.732 (0.504-0.864)	5.333	-16.13 to 13.46	30	12.806 (10.310)	11.630 (10.406)	0.629 (0.353-0.804)	6.309	-18.76 to 16.41
Upper arm NA	WDT (°C)	29	35.347 (1.323)	35.534 (1.798)	0.615 (0.327-0.799)	0.968	-2.53 to 2.90	30	35.312 (1.314)	35.594 (1.545)	0.337 (-0.018-0.618)	0.164	-2.95 to 3.52
	CDT (°C)	29	29.147 (1.270)	29.170 (1.257)	0.448 (0.104-0.696)	0.939	-2.61 to 2.60	30	29.081 (1.349)	28.536 (2.231)	0.412 (0.080-0.666)	1.372	-4.43 to 3.34
	HPT (°C)	29	42.232 (3.193)	42.810 (3.173)	0.791 (0.602-0.896)	1.455	-3.45 to 4.61	30	42.307 (3.164)	43.738 (2.894)	0.584 (0.244-0.786)	1.954	-3.60 to 6.46
	CPT (°C)	29	18.290 (8.694)	17.839 (8.833)	0.789 * (0.598-0.895)	4.025	-11.61 to 10.71	30	18.333 (8.546)	16.810 (8.869)	0.701 * (0.465-0.845)	4.761	-14.64 to 11.59
Trunk A	WDT (°C)	29	45.224 (5.232)	45.272 (5.88)	0.805 (0.626-0.904)	2.454	-6.76 to 6.86	30	45.383 (5.214)	44.741 (5.623)	0.707 (0.472-0.848)	2.933	-8.81 to 7.52
	CDT (°C)	29	14.437 (12.124)	13.625 (12.795)	0.876 (0.754-0.940)	4.387	-12.95 to 11.33	30	13.956 (12.201)	13.616 (13.862)	0.685 (0.433-0.837)	7.314	-20.90 to 20.22
	HPT (°C)	29	48.279 (2.555)	48.303 (2.584)	0.791 * (0.602-0.896)	1.175	-3.23 to 3.28	30	48.337 (2.530)	48.363 (3.324)	0.692 (0.445-0.841)	1.624	-4.57 to 4.62
	CPT (°C)	29	6.003 (9.466)	6.165 (9.404)	0.916 (0.829-0.960)	2.734	-7.41 to 7.74	30	5.803 (9.365)	6.038 (9.241)	0.722 (0.491-0.857)	4.905	-13.53 to 14.00

Trunk NA	WDT (°C)	29	35.716 (1.794)	35.608 (1.553)	0.529 (0.207-0.747)	1.148	-3.30 to 3.08	30	35.713 (1.763)	36.133 (1.373)	0.370 (0.027-0.638)	1.244	-3.04 to 3.88
	CDT (°C)	29	29.583 (1.098)	29.518 (1.029)	0.492 (0.159-0.724)	0.758	-2.17 to 2.04	30	29.564 (1.083)	29.294 (2.323)	0.333 (-0.028-0.616)	1.391	-4.38 to 3.84
	HPT (°C)	29	42.234 (3.691)	42.562 (3.373)	0.723 (0.490-0.860)	1.859	-3.45 to 4.61	30	42.330 (3.664)	43.404 (3.026)	0.625 (0.346-0.802)	2.048	-3.60 to 6.46
	CPT (°C)	29	17.818 (9.438)	17.787 (9.566)	0.874 * (0.749-0.939)	3.373	-11.61 to 10.71	30	18.132 (9.432)	17.072 (9.184)	0.678 * (0.426-0.832)	5.282	-14.64 to 11.59
Qceps	WDT (°C)	29	34.967 (1.740)	35.290 (1.512)	0.557 (0.245-0.764)	1.082	-2.69 to 3.33	30	34.942 (1.715)	35.621 (2.190)	0.361 (0.025-0.630)	1.561	-3.63 to 4.99
	CDT (°C)	29	29.237 (1.518)	28.777 (1.849)	0.630 (0.348-0.807)	1.024	-3.31 to 2.39	30	29.351 (1.618)	29.048 (1.944)	0.426 (0.085-0.678)	1.349	-4.06 to 3.46
	HPT (°C)	29	43.811 (3.265)	44.346 (2.889)	0.613 (0.324-0.798)	1.914	-4.78 to 5.85	30	43.831 (3.210)	45.124 (2.577)	0.602 (0.275-0.796)	1.825	-3.45 to 6.04
	CPT (°C)	29	15.083 (10.937)	16.958 (10.609)	0.768 (0.564-0.884)	5.189	-9.38 to 9.32	30	14.580 (11.094)	17.957 (9.864)	0.796 (0.545-0.906)	4.866	-15.77 to 13.65

Abbreviations: A = affected side, A1 = first assessment (rater 1), A2 = second assessment (rater 2), A3= third assessment (rater 1), CDT = cold detection thresholds, CI = Confidence Interval , CPT = cold pain thresholds, HPT = heat pain threshold, ICC = intraclass correlation coefficient, LoAs = Limits of agreement according to Bland-Altman, N = number of participants, NA = non-affected side, SD = standard deviation, SEM = standard error of measurement, Qceps = Quadriceps, WDT = warmth detection threshold.

* important influence of outlier on result (see Supplementary material S11 for results without outliers)

ICC > 0.90 = excellent reliability, 0.75-0.90 good reliability, 0.50-0.75 moderate reliability and < 0.50 weak reliability (12)

Table 6. *Intraclass correlations, means, standard error of measurement and limits of agreement for the evaluation of conditioned pain modulation*

	Intra-rater reliability						Inter-rater reliability					
	N	A1 Mean (SD)	A3 Mean (SD)	ICC (95% CI)	SEM	LoAs Lower to upper	N	A1 Mean (SD)	A2 Mean (SD)	ICC (95% CI)	SEM	LoAs Lower to upper
Test stimulus temp (°C)	27	41.59 (1.947)	42.11 (1.826)	0.877 (0.749-0.942)	0.661	-1.31 to 2.35	29	42.62 (1.879)	41.83 (1.947)	0.859 (0.724-0.931)	0.718	-1.78 to 2.19
Test 10s (NRS)	27	5.96 (1.829)	5.48 (2.137)	0.695 (0.434-0.848)	1.095	-3.52 to 2.56	29	6.00 (1.773)	5.83 (2.019)	0.687 * (0.434-0.840)	1.061	-3.13 to 2.79
Test 20s (NRS)	27	4.00 (1.922)	3.81 (1.777)	0.494 (0.148-0.733)	1.316	-3.83 to 3.46	29	4.00 (1.852)	3.66 (1.987)	0.399 (0.046-0.664)	1.488	-4.48 to 3.79
Test 30s (NRS)	27	3.41 (2.062)	2.85 (1.834)	0.491 (0.144-0.731)	1.390	-4.41 to 3.30	29	3.31 (2.020)	2.97 (2.079)	0.244 (-0.132-0.557)	1.782	-5.29 to 4.60
Test 40s (NRS)	27	2.74 (1.810)	2.41 (1.551)	0.472 (0.119-0.719)	1.221	-3.73 to 3.06	29	2.62 (1.801)	2.14 (1.552)	0.069 (-0.294-0.417)	1.618	-4.98 to 4.01
Conditioned test 10s (NRS)	27	4.56 (1.783)	4.52 (2.101)	0.357 (-0.020-0.645)	1.557	-4.37 to 4.29	29	4.52 (1.724)	4.79 (2.007)	0.444 (0.098-0.694)	1.310	-3.60 to 4.16
Conditioned test 20s (NRS)	27	3.48 (1.847)	3.67 (2.075)	0.531 (0.197-0.755)	1.343	-3.54 to 3.91	29	3.48 (1.785)	3.52 (1.920)	0.306 (-0.072-0.604)	1.543	-4.27 to 4.34
Conditioned test 30s (NRS)	27	2.93 (1.920)	2.81 (2.076)	0.497 (0.151-0.734)	1.417	-4.04 to 3.82	29	2.90 (1.858)	3.03 (1.700)	0.289 (-0.090-0.591)	1.500	-4.04 to 4.32
Conditioned test 40s (NRS)	27	2.63 (1.801)	2.44 (1.987)	0.540 (0.208-0.760)	1.284	-3.75 to 3.81	29	2.55 (1.764)	2.52 (1.299)	0.301 * (-0.078-0.600)	1.280	-3.64 to 3.57

CPM 10s	27	-1.41 (1.670)	-0.96 (1.605)	0.278 (-0.107-0.591)	1.391	-3.41 to 4.30	29	-1.48 (1.639)	-1.03 (1.700)	0.372 * (0.024-0.644)	1.317	-3.20 to 4.10
CPM 20s	27	-0.52 (2.225)	-0.15 (1.834)	0.268 (-0.0118-0.583)	1.736	-4.46 to 5.21	29	-0.52 (2.148)	-0.14 (1.382)	0.055 (-0.317-0.410	1.716	-4.49 to 5.25
CPM 30s	27	-0.48 (2.293)	-0.04 (1.720)	0.052 * (-0.329-0.417)	1.954	-5.03 to 5.92	29	-0.41 (2.228)	0.07 (1.307)	0.284 (-0.077-0.583)	1.496	-3.79 to 4.76
CPM 40s	27	-0.11 (1.805)	0.04 (1.765)	0.134 (-0.252-0.484)	1.661	-4.45 to 4.75	29	-0.07 (1.751)	0.38 (1.015)	0.030 (-0.326-0.383)	1.362	-3.45 to 4.35
CPM mean	27	-0.630 (1.783)	-0.278 (1.406)	0.176 (-0.212-0.516)	1.447	-2.87 to 1.96	29	-0.621 (1.724)	-0.181 (1.048)	0.264 (-0.094-0.566)	1.185	-2.63 to 1.83

Notes: Test = NRS score when test stimulus is applied without conditioning stimulus, Cond test = NRS score when test stimulus is applied together with conditioning stimulus, CPM = NRS score Cond test minus NRS score test (negative values indicate efficient conditioned pain modulation)

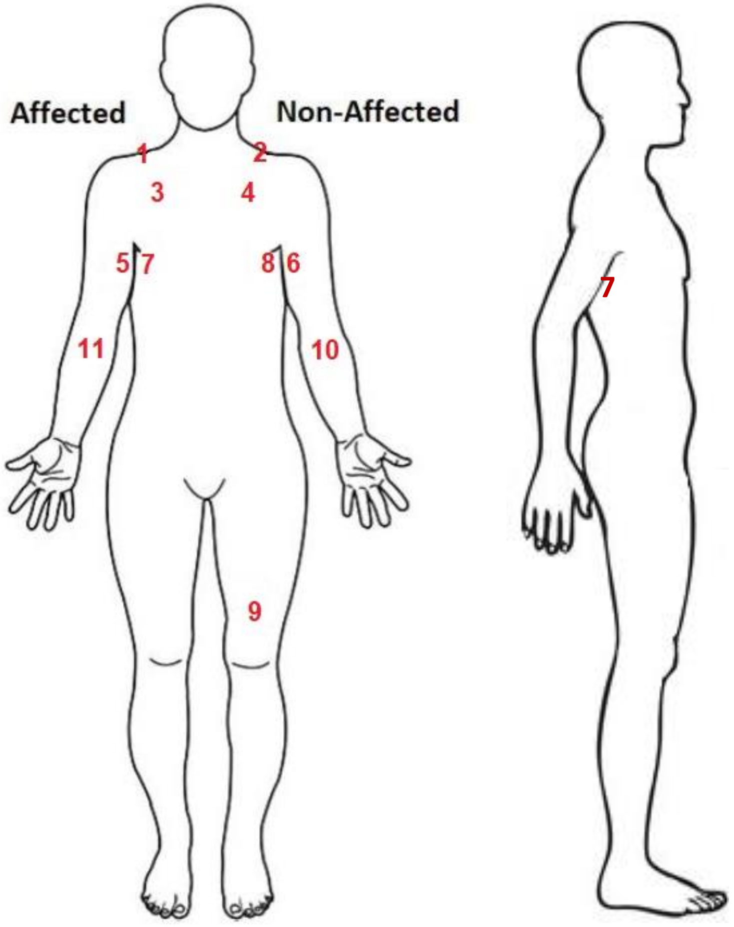
Abbreviations: A1 = first assessment (rater 1), A2 = second assessment (rater 2), A3 = third assessment (rater 1), CI = Confidence Interval, CPM = conditioned pain modulation, ICC = intraclass correlation coefficient, LoAs = Limits of agreement according to Bland-Altman, N = number of participants, NRS = numerical rating scale, s = seconds, SD = standard deviation, SEM = standard error of measurement.
* important influence of outlier on result (see Supplementary material S12 for results without outliers)
ICC > 0.90 = excellent reliability, 0.75-0.90 good reliability, 0.50-0.75 moderate reliability and < 0.50 weak reliability (12)

Table 7. *Intraclass correlations, means, standard error of measurement and limits of agreement for the evaluation of temporal summation*

	Intra-rater reliability						Inter-rater reliability					
	N	A1 Mean (SD)	A3 Mean (SD)	ICC (95% CI)	SEM	LoAs Lower to upper	N	A1 Mean (SD)	A2 Mean (SD)	ICC (95% CI)	SEM	LoAs Lower to upper
TS1 (NRS)	29	1.97 (1.918)	2.21 (1.989)	0.554 (0.241-0.763)	1.305	-3.37 to 3.86	30	1.90 (1.918)	2.47 (2.224)	0.688 (0.439-0.838)	1.157	-2.55 to 3.68
TS2 (NRS)	29	4.72 (2.359)	4.79 (2.381)	0.835 (0.679-0.919)	0.963	-2.60 to 2.74	30	4.60 (2.415)	4.83 (2.743)	0.883 (0.770-0.942)	0.882	-2.22 to 2.68
TS3 (NRS)	29	0.93 (1.462)	1.24 (1.883)	0.722 (0.489-0.859)	0.881	-2.15 to 2.77	30	0.90 (1.447)	1.17 (1.802)	0.802 (0.627-0.900)	0.773	-1.72 to 2.25
TS	29	2.76 (2.047)	2.59 (1.547)	0.620 (0.333-0.801)	1.108	-3.27 to 2.93	30	2.70 (2.037)	2.37 (1.956)	0.582 (0.290-0.776)	1.291	-3.91 to 3.24

Notes: TS1 = NRS score after first stimulation, TS2 = NRS score after 30s stimulation, TS3 = NRS score 15s after final stimulation, TS = NRS score T2 minus NRS score T1,

Abbreviations: A1 = first assessment (rater 1), A2 = second assessment (rater 2), A3 = third assessment (rater 1), CI = Confidence Interval, CPM = conditioned pain modulation, ICC = intraclass correlation coefficient, LoAs = Limits of agreement according to Bland-Altman, N = number of participants, NRS = numerical rating scale, s = seconds, SD = standard deviation, SEM = standard error of measurement, TS = temporal summation.
ICC > 0.90 = excellent reliability, 0.75-0.90 good reliability, 0.50-0.75 moderate reliability and < 0.50 weak reliability (12)



- Upper Trapezius **1-2**
  = muscle belly m. Trapezius (nn. supraclavicularis)

- Pectoral region **3-4**
  = index finger of the ipsilateral hand of the examiner at the height of the ipsilateral processus coracoidus of the patient, reference point under the ring finger at the m. Pectoralis Major
  (n. intercostalis medialis)

- Inner upper arm **5-6**
  = four fingers under the armpit fold at the height of the upper arm
  (n. intercostobrachialis)

- Lateral trunk **7-8**
  = four fingers under the armpit fold at the lateral trunk
  (n. intercostalis lateralis)

- Quadriceps **9**
  = wrist at the patellar base, reference point under the middle finger at the m. Quadriceps (n. femoralis)

- Lower arm **10-11**
  = middle volar side lower arm (n. cutaneus antebrachii medialis)