Scandinavian Journal of Pain 9 (2015) 38–41

Contents lists available at ScienceDirect

Scandinavian Journal of Pain

journal homepage: www.ScandinavianJournalPain.com

Observational study

Reliability of pressure pain threshold testing in healthy pain free young adults

Robert Waller[a,*], Leon Straker[a], Peter O'Sullivan[a], Michele Sterling[b], Anne Smith[a]

[a] Curtin University, GPO Box 1987, Perth, Western Australia 6845, Australia
[b] Griffith University Gold Coast Campus, Southport, Queensland 4222, Australia

HIGHLIGHTS

- Pressure pain threshold measurement is reliable using multiple research assistants.
- This study supports the use multiple research assistants in large cohort studies.
- It is recommended that raters are checked for systematic bias.
- Sample size calculations are provided for evaluating effects of interventions.

ARTICLE INFO

Article history:
Received 24 March 2015
Received in revised form 27 May 2015
Accepted 28 May 2015
Available online 15 July 2015

Keywords:
PPT (Pressure pain threshold)
Reliability
Multiple raters
Standard error of measurement

ABSTRACT

Background and aims: Investigation of the multidimensional correlates of pressure pain threshold (PPT) requires the study of large cohorts, and thus the use of multiple raters, for sufficient statistical power. Although PPT testing has previously been shown to be reliable, the reliability of multiple raters and investigation for systematic bias between raters has not been reported.

The aim of this study was to evaluate the intrarater and interrater reliability of PPT measurement by handheld algometer at the wrist, leg, cervical spine and lumbar spine. Additionally the study aimed to calculate sample sizes required for parallel and cross-over studies for various effect sizes accounting for measurement error.

Methods: Five research assistants (RAs) each tested 20 pain free subjects at the wrist, leg, cervical and lumbar spine. Intraclass correlation coefficient (ICC), standard error of measurement (SEM) and systematic bias were calculated.

Results: Both intrarater reliability (ICC = 0.81–0.99) and interrater reliability (ICC = 0.92–0.95) were excellent and intrarater SEM ranged from 79 to 100 kPa. There was systematic bias detected at three sites with no single rater tending to consistently rate higher or lower than others across all sites.

Conclusion: The excellent ICCs observed in this study support the utility of using multiple RAs in large cohort studies using standardised protocols, with the caveat that an absence of any confounding of study estimates by rater is checked, due to systematic rater bias identified in this study.

Implications: Thorough training of raters using PPT results in excellent interrater reliability. Clinical trials using PPT as an outcome measure should utilise a priori sample size calculations.

1. Introduction

Pressure pain threshold (PPT) measurement has been used extensively to investigate pain sensitivity (PS) in pain disorders [1–5]. Additionally, it is increasingly used as an outcome measure for evaluating interventions in musculoskeletal pain [6–11]. There are many potential biopsychosocial influences on PS that have not been comprehensively investigated to date and there is a lack of studies controlling for the multiple factors that potentially influence pain sensitivity. Large epidemiological cohorts provide an ideal opportunity to investigate PS where potential factors

DOI of refers to article: http://dx.doi.org/10.1016/j.sjpain.2015.06.002.
* Corresponding author at: School of Physiotherapy & Exercise Science, Curtin University, GPO Box 1987, Perth, Western Australia 6845, Australia. Tel.: +61 8 9266 3662; fax: +61 8 9266 3699.
E-mail addresses: R.Waller@curtin.edu.au (R. Waller), L.Straker@curtin.edu.au (L. Straker), P.OSullivan@curtin.edu.au (P. O'Sullivan), M.Sterling@griffith.edu.au (M. Sterling), Anne.Smith@exchange.curtin.edu.au (A. Smith).

http://dx.doi.org/10.1016/j.sjpain.2015.05.004
1877-8860/

contributing to PS can be examined concurrently. However the logistics of data collection in large cohort studies or randomised controlled trials often require multiple raters [12] raising the potential concern of interrater reliability.

Although previous interrater reliability studies for PPT using algometry have been conducted, these have varied with respect to sites assessed, number of raters examined, reliability statistics reported and degree of standardisation of algometry [1,13–15]. The most comphrehensive PPT reliability study to date reported excellent interrater reliability [1]. However, stated limitations of the study were the unknown applicability of the reliability estimates to raters with less training than the participating physiotherapists, and the inability to investigate potential systematic biases by rater.

The purpose of the current study was to assess the reliability of PPT measurement by research assistants (RAs), who were employed to take a range of measures including PPT in a large cohort study of healthy young adults. The study aimed to assess the intrarater and interrater reliability, including systematic bias, of PPT testing by the same method (handheld algometer) and at the same body sites (wrist, leg, cervical and lumbar spine) as in the larger cohort study. Additionally the study aimed to calculate sample sizes required for parallel and cross-over studies for various effect sizes accounting for measurement error.

2. Methods

2.1. Subjects

The raters were five female RAs collecting data for the Western Australian Pregnancy Cohort (Raine) study. The mean (SD) age, weight and height respectively were 37 years [10], 64.4 (5.7) kg and 167 (9) cm. Four raters had obtained a Bachelor of Science degree and one a Bachelor of Arts degree. The RAs had on average 5.4 years' experience of data collection, performing a wide range of physical measures and tests.

A convenience sample of 20 pain-free young adults were recruited from students of the Faculty of Health, Curtin University as test subjects. Subjects were included on the basis that they had not reported wrist, lower leg, neck or low back pain in the previous 3 months. Subjects were excluded if they did not perceive pressure below 1000 kPa during PPT testing. All participants signed an informed consent form prior to testing and basic demographic data consisting of age, weight, sex and height was obtained.

2.2. Study protocols

2.2.1. Phase 1: rater training

The raters were formally trained in subject instruction and pressure algometer application for a total of 3 h over 3 occasions. The algometer (Somedic AB, Sweden) used had a circular contact area of 1 cm^2. Raters were trained in standardising their rate of pressure application, accurate land marking of test sites, applying pressure perpendicular to the skin and correct handling of the algometer to achieve effectiveness of pressure application, particularly for when a subject had a PPT near the 1000 kPa cut off. Raters were tested in their consistency of pressure application and over five consecutive applications were required to achieve, a rate of 50 kPa/s over a ten second period. The rate of pressure application was considered acceptable if force was applied at a rate of 50 kPa/s +/−10 kPa/s, resulting in a pressure reading of 500 kPa +/−100 kPa after 10 s. This threshold of variation in the rate of pressure application has been used in previous reliability studies [1,13].

2.2.2. Phase 2: reliability testing

Four sites were tested in the following sequence; the dorsal wrist, upper leg, cervical spine and lumbar spine. The wrist was tested at the middle of the dorsal aspect of the wrist joint line. The leg was tested on the muscle belly of tibialis anterior, approximately 2.5 cm lateral and 5 cm distal to the tibial tubercle. The neck was tested on the trapezius muscle, at the mid-point between the C7 spinous process and the lateral acromion. The lumbar spine was tested at the erector spinae, 2 cm lateral to the L4/L5 interspinous space. The right side was used for all subjects. The test sites were identified by each rater prior to testing and participants were positioned in a standardised manner. The algometer was applied perpendicularly to the skin at each of the four sites tested.

PPT was defined as the moment pressure increased to a point where it first felt uncomfortable or painful. Prior to testing standardised instructions read to participants included, "Pressure will be applied at a gradual rate. Allow the pressure to increase until it reaches a point where it first feels uncomfortable or painful and then press the button. This means the very first onset of discomfort or pain and not the most pressure that you can bear." The pressure started at 0 kPa and increased at a constant rate of 50 kPa/s until pain threshold was reached and the participant terminated the test by pressing a hand held switch. A maximum of 1000 kPa was set for safety purposes.

Testing was performed on one day in a large, temperature-controlled room, with five raters positioned at stations separated by curtains. Subjects rotated between the five raters according to a pre-specified randomised order, with a 10-min rest between stations. Subjects were blinded to their PPT test values, and raters were blinded to other raters test values. The rater who tested a subject first performed three tests at each site (the first allowed familiarisation with the testing procedure and was not used for analysis) and two tests on subsequent subjects. A 10-s rest was given between tests at the same site.

2.3. Statistical analysis

A-priori power calculation indicated a sample size of 20 participants and two tests per tester would provide over 90% power to detect a standard error of measurement exceeding an acceptable limit of 100 kPa [16]. Data was analysed using IBM SPSS Statistics 21.0 (Chicago, USA). Relative estimates of interrater and intrarater reliability (ICCs) were calculated using an ICC for absolute agreement estimated under a two-way random effects model, using both measures taken by each rater for intrarater ICCs and the average of the two measures taken by each rater for interrater ICCs. Intrarater ICCs were calculated for each rater. Standard error of measurement (SEM) was used as an absolute estimate of interrater reliability and was calculated for each site as the square root of the mean square error term of the repeated measures analysis of variance test [17], which also assessed systematic differences between raters. Lastly, sample sizes required for parallel and cross-over studies to detect changes or differences from baseline for each testing site adjusted for measurement error were calculated. A range of change or difference from 10 to 30% using the mean of the sample as baseline was adjusted for measurement error using Guyatt's responsiveness index [18] (change divided by the square root of two times the mean squared error term from the repeated measures analysis of variance test). G*Power 3.1 [19] was used for sample size calculations.

3. Results

3.1. Subjects

20 pain free subjects (50% female) between the ages of 20 and 33 years old were tested. The mean (SD) age, weight and height

Table 1
Interrater reliability.

Site	ICC absolute (95% CI)	SEM (kPa)
Wrist	0.93 (0.87–0.97)	84
Leg	0.92 (0.84–0.97)	93
Neck	0.92 (0.85–0.97)	79
Back	0.95 (0.90–0.98)	100

Abbreviations: ICC, intraclass correlation coefficient; CI, confidence interval; SEM, standard error of measurement.

respectively were 23.3 (3.8) years, 73 (13.3) kg and 175 (11) cm. The mean (SD) values for PPT at the wrist, leg, neck and back were respectively 472 kPa (146), 501 kPa (183), 413 kPa (133) and 606 kPa (224).

3.2. Intrarater reliability

Excellent values for intrarater reliability were obtained at all sites. ICC (absolute) values for the five raters ranged from 0.81 to 0.97 at the wrist, 0.96 to 0.98 at the leg, 0.92 to 0.98 at the neck and 0.94 to 0.99 at the back.

3.3. Interrater reliability

Excellent ICC and SEM values for interrater reliability were also obtained at all sites (Table 1). Systematic differences were detected between raters at the back ($p = .003$), neck ($p = .013$) and leg ($p < .001$) testing sites but not at the wrist ($p = .252$). The maximum interrater difference observed was 190 kPa at the leg, 79 kPa at the neck and 129 kPa at the back 181, with no single rater tending to consistently rate higher or lower than others across all sites.

3.4. Sample size calculation

Table 2 presents sample size requirements for a 10–30% change in PPT at each test site accounting for measurement error using means of the sample as initial mean values for percentage change calculations.

4. Discussion

The results of this study demonstrate high intrarater reliability of RAs when measuring PPTs over the four sites. This study adds to existing data and the results compare favourably to previous investigations in pain free subjects. Walton et al. [1] and Persson et al. [20] found intrarater ICCs of 0.97 and 0.70–0.90 respectively for PPT measurement at the upper trapezius, comparable to the range of ICCs for the five raters of 0.92–0.98 in the current study. Walton et al. [1] also tested PPT reliability at the tibialis anterior and found an intrarater ICC of 0.94, again comparable to the result of 0.96–0.98 for the raters in this study.

The results of this study also demonstrate high interrater reliability as measured by relative measures (ICCs). The interrater ICCs of 0.95 (95% CI: 0.90–0.98) at the leg and 0.92 (95% CI: 0.85–0.97) at the neck compare favourably to previous reports of Walton et al. [1] at these sites for pain free individuals of 0.84 (95% CI: 0.75–0.90) at the leg and 0.79 (95% CI: 0.66–0.87) at the neck. Reliability of PPT measurement at the lumbar spine and dorsal wrist have not been reported previously, and this study confirms high interrater reliability can be achieved at these sites. Other previous reliability studies for PPT have shown high levels of interrater reliability using only two raters [1,14,21–23] at sites in the upper quarter in pain-free individuals. The current study evaluated raters who were RAs rather than physiotherapists used by Walton et al. [1] and Nussbaum et al. [14], and provides evidence that with systematic training excellent measures of relative reliability can be achieved by raters of differing backgrounds.

Although interrater ICCs, which are based on rankings, were excellent, these do not highlight the magnitude of between rater variation in measures or any systematic bias. Inter-rater SEM results of 79 kPa at the neck and 93 kPa at the leg were somewhat higher than Walton et al. [1] who reported 53 kPa and 59 kPa at the neck and leg respectively. We also detected systematic bias, and this might be partly the reason for the higher SEMs in this study. This finding is important with regard to studies seeking to evaluate associations between PPT and other factors, or differences in PPT between pain/disorder groups, using more than one rater. Systematic bias of raters, means that there is the potential that measures of association or difference will be confounded by rater. Therefore, for

Table 2
Sample size calculations.

Site (initial mean value)	% change	SEM (kPa)	ES	Sample Size required for crossover design		Sample size required for parallel design	
				80% power	90% power	80% power	90% power
Wrist (472 kPa)	30	84	1.19	8	9	10	11
	25	84	0.99	11	12	13	14
	20	84	0.79	15	16	17	18
	15	84	0.60	24	25	26	27
	10	84	0.40	52	53	54	55
Leg (501 kPa)	30	93	1.14	9	10	11	12
	25	93	0.95	11	12	13	14
	20	93	0.76	16	17	18	19
	15	93	0.57	27	28	29	30
	10	93	0.38	57	58	59	60
Neck (413 kPa)	30	79	1.11	9	10	11	12
	25	79	0.92	12	13	14	15
	20	79	0.74	17	18	19	20
	15	79	0.55	28	29	30	31
	10	79	0.37	60	61	62	63
Back (606 kPa)	30	100	1.29	7	8	9	10
	25	100	1.07	9	10	11	12
	20	100	0.86	13	14	15	16
	15	100	0.64	22	23	24	25
	10	100	0.43	45	46	47	48

Abbreviations: SEM, standard error of measurement; ES, effect size (Guyatt's responsiveness index = percentage change divided by the square root of two times the mean square error term from the repeated measures analysis of variance test).

studies of associations or differences between PPT and other factors using multiple raters, it is recommended that a sensitivity analysis be performed by adjusting estimates for rater, to avoid confounding of group differences or associations by rater bias.

PPT is increasingly being used as an outcome measure in clinical trials for musculoskeleta pain [6–9]. However, only a few studies report a priori sample size calculation [6,10,11]. For sample size calculations it is advantageous to account for spurious change (i.e. measurement error) to achive sufficient power to detect minimal clinically important changes or differences that exceeds measurement error. Table 2 presents a guide for sample size calculations according to percentage change, accounting for measurement error.

5. Conclusion

In conclusion, this study established the reliability of PPT measurement using handheld algometry by multiple RAs at multiple body sites. Large studies with sufficient power to identify the many likely multivariable correlates of pain sensitivity are needed, and this often necessitates the use of multiple raters. The sample size calculations presented will assist researchers to determine sample sizes which account for measurement error for interventions using PPT as an outcome measure. The results of this study support the utility of using multiple RAs in large cohort studies using standardised protocols, with the caveat that an absence of any confounding of study estimates due to potential systematic rater bias be checked.

Ethical statement

All aspects of the study were approved by Curtin University Human Research Ethics Committee.

Conflict of interest

The authors have no conflict of interest.

References

[1] Walton DM, Macdermid JC, Nielson W, Teasell RW, Reese H, Levesque L. Reliability, standard error, and minimum detectable change of clinical pressure pain threshold testing in people with and without acute neck pain. J Orthop Sports Phys Ther 2011;41:644–50.
[2] Slater H, Arendtnielsen L, Wright A, Gravennielsen T. Sensory and motor effects of experimental muscle pain in patients with lateral epicondylalgia and controls with delayed onset muscle soreness. Pain 2005;114:118–30.
[3] Zhou Q, Fillingim RB, Riley Iii JL, Malarkey WB, Verne GN. Central and peripheral hypersensitivity in the irritable bowel syndrome. Pain 2010;148:454–61.
[4] Suokas AK, Walsh DA, McWilliams DF, Condon L, Moreton B, Wylde V, Arendt-Nielsen L, Zhang W. Quantitative sensory testing in painful osteoarthritis: a systematic review and meta-analysis. Osteoarthr Cartil 2012;20:1075–85.
[5] O'Neill S, Manniche C, Graven-Nielsen T, Arendt-Nielsen L. Generalized deep-tissue hyperalgesia in patients with chronic low-back pain. Eur J Pain 2007;11:415–20.
[6] Kardouni JR, Shaffer S.W., Pidcoe P.E., Finucane S.D., Cheatham S.A., Michener L.A. Immediate changes in pressure pain sensitivity after thoracic spinal manipulative therapy in patients with subacromial impingement syndrome: a randomized controlled study. Manual Therapy 2015;20:540–6.
[7] Macedo LB, Josué AM, Maia PHB, Câmara AE, Brasileiro JS. Effect of burst TENS and conventional TENS combined with cryotherapy on pressure pain threshold: randomised, controlled, clinical trial. Physiotherapy 2015;101:155–60.
[8] Ylinen J, Takala E-P, Kautiainen H, Nykänen M, Häkkinen A, Pohjolainen T, Karppi S-L, Airaksinen O. Effect of long-term neck muscle training on pressure pain threshold: a randomized controlled trial. Eur J Pain 2005;9:673–81.
[9] Bakar Y, Sertel M, Öztürk A, Yümin ET, Tatarli N, Ankarali H. Short term effects of classic massage compared to connective tissue massage on pressure pain threshold and muscle relaxation response in women with chronic neck pain: a preliminary study. J Manipulative Physiol Ther 2014;37:415–21.
[10] Fuentes CJ, Armijo-Olivo S, Magee DJ, Gross DP. A preliminary investigation into the effects of active interferential current therapy and placebo on pressure pain sensitivity: a random crossover placebo controlled study. Physiotherapy 2011;97:291–301.
[11] Venancio RC, Pelegrini S, Gomes DQ, Nakano EY, Liebano RE. Effects of carrier frequency of interferential current on pressure pain threshold and sensory comfort in humans. Arch Phys Med Rehabil 2013;94:95–102.
[12] Rolke R, Baron R, Maier C, Tolle TR, Treede RD, Beyer A, Binder A, Birbaumer N, Birklein F, Botefur IC, Braune S, Flor H, Huge V, Klug R, Landwehrmeyer GB, Magerl W, Maihofner C, Rolko C, Schaub C, Scherens A, Sprenger T, Valet M, Wasserka B. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): standardized protocol and reference values. Pain 2006;123:231–43.
[13] Chesterton LS, Sim J, Wright CC, Foster NE. Interrater reliability of algometry in measuring pressure pain thresholds in healthy humans, using multiple raters. Clin J Pain 2007;23:760–6.
[14] Nussbaum EL, Downes L. Reliability of clinical pressure-pain algometric measurements obtained on consecutive days. Phys Ther 1998;78:160–9.
[15] Tunks E, McCain GA, Hart LE, Teasell RW, Goldsmith CH, Rollman GB, McDermid AJ, DeShance PJ. The Reliability or examination for tenderness in patients with myofascial pain, chronic fibromyalgia and controls. J Rheumatol 1995;22:944–52.
[16] Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. Phys Ther 1997;77:745–50.
[17] Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res 2005;19:231–40.
[18] Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40:171–8.
[19] Faul F, Erdfelder E, Buchner A, Lang A-G. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009;41:1149–60.
[20] Persson AL, Brogardh C, Sjolund BH. Tender or not tender: test–retest repeatability of pressure pain thresholds in the trapezius and deltoid muscles of healthy women. J Rehab Med 2004;36:17–27.
[21] Geber C, Klein T, Azad S, Birklein F, Gierthmühlen J, Huge V, Lauchart M, Nitzsche D, Stengel M, Valet M, Baron R, Maier C, Tölle T, Treede R-D. Test–retest and interobserver reliability of quantitative sensory testing according to the protocol of the German Research Network on Neuropathic Pain (DFNS): a multicentre study. Pain 2011;152:548–56.
[22] Delaney GAM, McKee ACM. Inter and intra-rate reliability of the pressure pain threshold meter in measurement of myofascial trigger point sensitivity. Am J Phys Med Rehabil 1993;72:136–9.
[23] Reeves JL, Jaeger B, Graff-Radford SB. Reliability of the pressure algometer as a measure of myofascial trigger point sensitivity. Pain 1986;24:313–21.