Methodische Herausforderungen der Messung chronischer Schmerzen bei der Durchführung klinischer Studien

Dissertation

zur Erlangung des akademischen Grades

doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Fakultät für Sozial- und Verhaltenswissenschaften
der Friedrich-Schiller-Universität Jena

von M.A. Sportwiss. Helen Nothnagel

geboren am 10. November 1987 in Halberstadt

Gutachter

1. PD Dr. Christian Puta
   Friedrich-Schiller-Universität Jena, Institut für Sportwissenschaft, Sportmedizin und Gesundheitsförderung

2. Prof. Dr. Heiko Wagner
   Westfälische Wilhelms-Universität Münster, Institut für Sportwissenschaft, Bewegungswissenschaft

Tag der mündlichen Prüfung: 27. Oktober 2021

Zusammenfassung

Die Entstehung von Schmerz sowie die Weiterleitung der Schmerzreize über das Nervensystem (Nozizeption) sind komplexe Prozesse, die stets gleich ablaufen. Das Schmerzempfinden und die Schmerzwahrnehmung sind dagegen subjektiv und hängen unter anderem von geschlechtsbezogenen, soziokulturellen und gesellschaftlichen Faktoren ab. Während beim akuten Schmerz die Ursachenfindung und gezielte Therapie im Vordergrund stehen, ist beim chronischen Schmerz die ursprüngliche Ursache oftmals nicht mehr zu finden. Die Komplexität der chronischen Schmerzen ist sowohl in der klinischen Praxis als auch in der Schmerzforschung sehr herausfordernd. Darüber hinaus erschwert auch die subjektive Komponente der Schmerzwahrnehmung die Schmerzdiagnostik, vor allem die Messung der Schmerzstärke und der Schmerzempfindlichkeit.

Die vorliegende Dissertation beschäftigt sich mit den Herausforderungen bei der Durchführung klinischer Studien zum chronischen Schmerz, die auf der Komplexität chronischer Schmerzen, deren Diagnostik und Erfassung basieren. Zunächst wird der aktuelle Forschungsstand zum Schmerz bzw. zum chronischem Schmerz aufgezeigt. Darauf aufbauend werden drei Untersuchungen vorgestellt, die sich mit methodenkritischen Aspekten bei der Durchführung von klinischen Studien zum chronischen Schmerz auseinandersetzen.

Die erste Untersuchung (Nothnagel et al., 2019) beschäftigt sich mit dem zeitkritischen Faktor der Rekrutierung von chronischen Schmerzpatienten in klinischen Studien. Sie untersucht die signifikante Abnahme der Schmerzintensität während der Rekrutierungsphase einer klinischen Schmerzstudie, das heißt zwischen der ersten Kontaktaufnahme mit dem potenziellen Patienten am Telefon und der ersten Untersuchung. Es werden drei Haupteinflüsse identifiziert: der natürliche Krankheitsverlauf, der Hawthorne Effekt und das statistische Phänomen der Regression zum Mittelwert. Darüber hinaus werden Empfehlungen für Rekrutierungsverfahren in klinischen Studien zum chronischem Schmerz gegeben. Demzufolge können das Einschlusskriterium der Schmerzintensität sowie die Zeitspanne der gesamten Rekrutierungsphase entsprechend angepasst werden. Des Weiteren scheint eine regelmäßige und engmaschige Dokumentation der Schmerzintensität während

der gesamten Rekrutierung hilfreich, um möglicherweise Symptomveränderungen und eben auch den natürlichen Krankheitsverlauf kontrollieren zu können.

In der zweiten Untersuchung (Nothnagel et al., 2017) wird erstmalig die Zuverlässigkeit (Test-Retest-Reliabilität) und Übereinstimmung des gesamten Protokolls der standardisierten Quantitativen Sensorischen Testung (QST) bei gesunden Probanden über einen therapierelevanten Zeitraum von 10 Wochen analysiert. Die meisten QST-Parameter sind über den untersuchten Zeitraum reliabel. Darüber hinaus identifizieren die Ergebnisse Übereinstimmungsgrenzen (Limits of agreement) für alle QST-Parameter. Die Kenntnis über die Variabilität der QST-Parameter ist eine notwendige Voraussetzung, um die durch Interventionen induzierten Veränderungen der Parameter beurteilen zu können. Die Ergebnisse der Studie legen nahe, dass die Methode der QST grundsätzlich sehr robust ist. Das heißt aber auch, dass die QST nicht sensitiv genug ist, um kleine Veränderungen über die Zeit erkennen zu können. Veränderungen über die Zeit müssen außerhalb der Übereinstimmungsgrenzen liegen, um überhaupt als solche erkannt werden zu können.

Abschließend werden die erhobenen Daten der beiden zuvor genannten Studien in einer dritten Analyse verglichen und die QST-Profile der mitteleuropäischen Studienpopulation der subarktischen Studienpopulation gegenübergestellt (Nothnagel et al., 2021). Die Analyse zeigt einen Unterschied der thermischen Schmerzschwellen des QST-Protokolls (Kälteschmerzschwelle und Hitzeschmerzschwelle) zwischen den beiden untersuchten Populationen. Im Vergleich zur mitteleuropäischen Stichprobe zeigt die subarktische Studienpopulation eine höhere Schmerzempfindlichkeit sowohl gegenüber Kälte als auch gegenüber Hitze. Im Gegensatz dazu zeigen die mechanischen Schwellen sowie die thermischen Detektionsschwellen keine Unterschiede zwischen den Populationen. Nach bisherigem Kenntnisstand finden sich in der Literatur keine vergleichbaren Publikationen, die diesen Ansatz untersucht haben. Um klimabedingte Unterschiede der thermischen Schmerzschwellen weiter aufzuklären und das Ausmaß dieses Effekts in klinischen Studien zum chronischen Schmerz abschätzen zu können, ist es dringend erforderlich, den hier gezeigten Effekt systematisch weiter zu untersuchen.

Die vorliegende Arbeit verfügt über eine große Allgemeinrelevanz. Die referierten Ergebnisse haben eine umfassende Bedeutung für klinische Studien zum chronischen Schmerz. Es werden Empfehlungen gegeben, die sowohl in der klinischen Praxis als auch in der Forschung zum chronischen Schmerz anwendbar sind.

Abstract

The development of pain and the transmission of pain stimuli via the nervous system (nociception) are complex processes that usually occur in the same way. In contrast, sensation as well as perception of pain are subjective and depend, among other factors, on gender-related, sociocultural, and social factors. While the focus is on finding the cause and targeted therapy in acute pain, the original cause is often no longer identifiable in chronic pain. The complexity of chronic pain is challenging both in clinical practice and in pain research. In addition, the subjectivity of pain perception complicates pain diagnostics, especially with regard to the measurement of pain intensity and pain sensitivity.

The present thesis addresses the challenges of conducting clinical trials on chronic pain based on the complexity of chronic pain, its diagnosis and recording. Firstly, the current state of research on pain and chronic pain is presented. Based on this, three studies are presented that deal with method-critical aspects in the conduct of clinical studies on chronic pain.

The first study (Nothnagel et al., 2019) investigates the time-critical factor of the recruitment of chronic pain patients in clinical trials. It examines the significant decrease in pain intensity during the recruitment of the clinical pain study, i.e. between the first contact with the potential patient on the phone and the first examination. Three main influences were identified: the natural history of disease, the Hawthorne effect, and the regression-to-the-mean phenomenon. Recommendations for recruitment procedures for clinical trials on chronic pain are given. Accordingly, the inclusion criterion of pain intensity as well as the time interval of the entire recruitment period should be adjusted. Furthermore, a close documentation of the pain intensity during the whole recruitment period seems to be helpful in order to monitor possible symptom changes and the natural course of the disease.

The second study (Nothnagel et al., 2017) analyzes the reliability (test-retest reliability) and limits of agreement of the entire protocol of the standardized quantitative sensory testing (QST) in healthy subjects over a therapy-relevant period of 10 weeks for the first time. Most QST parameters are reliable over this time period. In addition, the results identify limits of agreement for all QST parameters. The knowledge about the variability of QST parameters is a necessary prerequisite to assess changes of the parameters induced by interventions. The results of the study suggest that the method of QST is in principal very robust. However, this

also means that the QST may not be sensitive enough to detect small changes over time, since these need to be outside the limits of agreement to be detected.

Finally, the data from the two studies were compared in a third analysis (Nothnagel et al., 2021). The analysis shows a difference in the thermal pain thresholds of the QST protocol (cold pain threshold and heat pain threshold) between the two study populations. Compared to the mid-European population, the sub-Arctic population shows higher pain sensitivity to cold and heat. In contrast, mechanical thresholds as well as thermal detection thresholds show no differences between the populations. Based on current state of knowledge, there are no comparable publications in the literature that have investigated this approach. There is a need to further systematically investigate the effect shown here further in order to elucidate climate-related differences in thermal pain thresholds and to estimate the magnitude of the effect in clinical studies of chronic pain.

The results presented in this thesis are of high relevance and have broad significance for clinical studies of chronic pain. Recommendations are given which are applicable in clinical practice as well as in clinical research on chronic pain.

Inhaltsverzeichnis

Abkürzungsverzeichnis

CRPS	Komplexes regionales Schmerzsyndrom (engl. complex regional pain syndrome)
CDT	Kältedetektionsschwelle (engl. cold detection threshold)
CPT	Kälteschmerzschwelle (engl. cold pain threshold)
DFNS	Deutscher Forschungsverbund Neuropathischer Schmerz (engl. German Research Network on Neuropathic Pain)
DMA	Dynamisch mechanische Allodynie (engl. dynamic mechanical allodynia)
HPT	Hitzeschmerzschwelle (engl. heat pain threshold)
IASP	Internationalen Gesellschaft zum Studium des Schmerzes (engl. International Association for the Study of Pain)
ICC	Intraklassen-Korrelationskoeffizient (engl. intraclass correlation coefficient)
ICD	Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (engl. International Statistical Classification of Diseases and Related Health Problems)
IF	Impact-Faktor (engl. impact factor)
LoA	Übereinstimmungsgrenzen (engl. limits of agreement)
MDT	Mechanische Detektionsschwelle (engl. mechanical detection threshold)
MPS	Mechanische Schmerzschwelle (engl. mechanical pain sensory)
MPT	Mechanische Schmerzschwelle (engl. mechanical pain threshold)
NRS	Nummerische Ratingskala (engl. numeric rating scale)
PPT	Druckschmerzschwelle (engl. pressure pain threshold)
PHS	Paradoxe Hitzeempfindung (engl. paradoxical heat sensation)
QST	Quantitative Sensorische Testung (engl. quantitative sensory testing)
SEM	Standardfehler der Messung (engl. standard error of measurement)
TSL	Thermische Unterschiedsschwelle (engl. thermal sensory limen)
VAS	Visuelle Analogskala (engl. visual analogue scale)
VDT	Vibrationsdetektionsschwelle (engl. vibration detection threshold)
VRS	Verbale Ratingskala (engl. verbale rating scale)
WDT	Wärmedetektionsschwelle (engl. warm detection threshold)
WUR	Wind-up (engl. wind-up ratio)

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Bereits Mitte des 17. Jahrhunderts formulierte der französische Philosoph und Naturwissenschaftler René Descartes (1596 - 1650) ein erstes Verständnis von Schmerz. Er beschrieb in seiner „Abhandlung über den Menschen“ mithilfe seines „Glockenstrang-Schmerzmodells“ Schmerzen als direkte Folge einer körperlichen Schädigung. Diese Schädigung löst einen Nervenreiz aus, welcher über Nervenbahnen ins Gehirn geleitet werde, als zöge man am Seil einer Glocke (Abbildung 1). Das Gehirn gibt dann den Befehl zu handeln, um infolgedessen zum Beispiel den Fuß vom Feuer zurückzuziehen, aufzustehen oder sich zu schonen (Egloff et al., 2008).

Abb. 1. Weg des Schmerzes nach R. Descartes (Egloff et al., 2008, S. 550)

Im Laufe der Schmerzforschung bildeten sich weitere Schmerztheorien heraus. Hierzu zählt zum Beispiel die „Gate-Control-Theorie“ von Melzak und Wall (Melzak und Wall, 1965; Moayedi und Davis, 2013). Mitte der Siebzigerjahre des 20. Jahrhunderts wurde von der Internationalen Gesellschaft zum Studium des Schmerzes (IASP) eine Schmerzdefinition erstellt, die der Komplexität der Schmerzverarbeitung und des Schmerzempfindens gerecht wird: *„Pain is an unpleasant sensory and emotional experience with actual or potential tissue damage or described in terms of such damage”* (Schulte am Esch & Bause, 2011, S. 593).

Klassischerweise werden Schmerzen in akut und chronisch unterteilt. Nicht selten entwickeln sich chronische Schmerzen über einen längeren Zeitraum, werden allmählich stärker und breiten sich nach und nach im Körper aus. Dann ist der ursprüngliche Auslöser oft gar nicht mehr zu identifizieren. Chronische Schmerzdiagnosen wurden bisher nicht systematisch dargestellt. Erst in der elften Version der Internationalen statistischen Klassifikation von Krankheiten und verwandten Gesundheitsproblemen (ICD-11), die zum 1. Januar 2022 in Kraft treten wird, werden chronische Schmerzen nicht mehr nur als Symptom, sondern als eigenständiges Krankheitsbild beurteilt. Unter der Diagnose-Ziffer MG30 findet sich eine eigenständige Kategorie für die Klassifikation chronischer Schmerzen. Diese umfasst sieben Kategorien chronischer Schmerzen (chronisch primäre Schmerzen, chronischer Krebsschmerz, chronisch postchirurgischer Schmerz, chronisch-neuropathischer Schmerz, chronischer Kopfschmerz, chronisch viszeraler Schmerz und chronisch muskuloskelettaler Schmerz) und unterstreicht einmal mehr die Komplexität chronischer Schmerzen (Treede et al., 2015).

Nicht zuletzt erschwert diese Komplexität der chronischen Schmerzen die Erfassung und Diagnostik chronischer Schmerzen in klinischen Studien. Sowohl in der Forschung als auch in der klinischen Praxis werden Schmerzskalen eingesetzt, die Intensität von Schmerzen zu erfassen. Um die Schmerzempfindlichkeit zu messen, wird oftmals die Methode der Quantitativ Sensorischen Testung (QST) angewandt. Doch auch mit der QST wird deutlich, wie die individuell sehr unterschiedliche Schmerzwahrnehmung und -empfindung eine objektive Schmerzmessung substanziell erschwert.

Die vorliegende Dissertationsschrift analysiert methodische Herausforderungen der Messung chronischer Schmerzen bei der Durchführung klinischer Studien. Dabei werden anhand von drei auf eigenen Untersuchungen basierenden Publikationen methodenkritische Aspekte bei der Durchführung klinischer Studien zum chronischen Schmerz diskutiert.

Die wissenschaftliche Qualität und Aussagekraft einer klinischen Studie wird maßgeblich durch die präzise Studienplanung und deren Umsetzung im Studienverlauf bestimmt. Die passende Auswahl des Studiendesigns (**experimentelle und quasi-experimentelle Studiendesigns**) ist dabei essenziell. Ein Studiendesign in der klinischen Forschung bezieht hauptsächlich die Aspekte der Fragestellung, Studienpopulation, Studientyp, Beobachtungseinheit, Messverfahren und Fallzahlplanung ein (Röhrig et al., 2009; Abbildung 2). Die Auswahl der geeigneten Studienpopulation umfasst die Rekrutierung

(beispielsweise Art, Gebiet und Zeit), soziodemografische Angaben zu den Probanden (beispielsweise Alter, Geschlecht und Krankheit) sowie Ein- und Ausschlusskriterien. Treten während der Rekrutierungsphase, also noch vor Beginn der Studie bzw. Intervention, Veränderungen relevanter Outcome-Parameter auf, kann das zu erheblichen Problemen bei der Studienauswertung führen. Darüber hinaus ist für den Erfolg einer klinischen Studie die richtige Auswahl der Beobachtungseinheit bzw. des Biomarkers entscheidend. Dabei ist es wichtig, einen Biomarker auszuwählen, der exakt die studienrelevanten Parameter abbilden kann. Zugleich muss der ausgewählte Biomarker so exakt wie möglich den zu untersuchenden Parameter messen. Und vor allem muss der entsprechende Biomarker so änderungssensitiv sein, dass eben auch die Veränderungen, die man sehen will, durch den Biomarker abgebildet werden können. Diese Aspekte finden sich unter dem Schwerpunkt der Testgütekriterien (**Gütekriterien**), speziell der Reliabilität, wieder. Neben der Reliabilität zählen auch die Objektivität und die Validität zu den Hauptgütekriterien (Röhrig et al., 2009). Eine gut ausgewählte Studienpopulation lässt die Übertragbarkeit der mit der klinischen Studie analysierten Ergebnisse auf die Zielpopulation zu. Es stellt sich also die Frage nach der Generalisierbarkeit der Studienergebnisse. Darüber hinaus ist entscheidend, wie relevant die Studienergebnisse für die weitere Forschung bzw. für die klinische Praxis sind. Dieser Aspekt umfasst die **ökologische Validität**.

Abbildung 2 zeigt eine Übersicht der relevanten Aspekte der experimentellen und quasi-experimentellen Studiendesigns, der Gütekriterien und der ökologischen Validität sowie deren Interaktion. Die hervorgehobenen (grau markierten) Aspekte stellen die Hauptuntersuchungsgegenstände der in dieser Arbeit vorgestellten Publikationen dar.

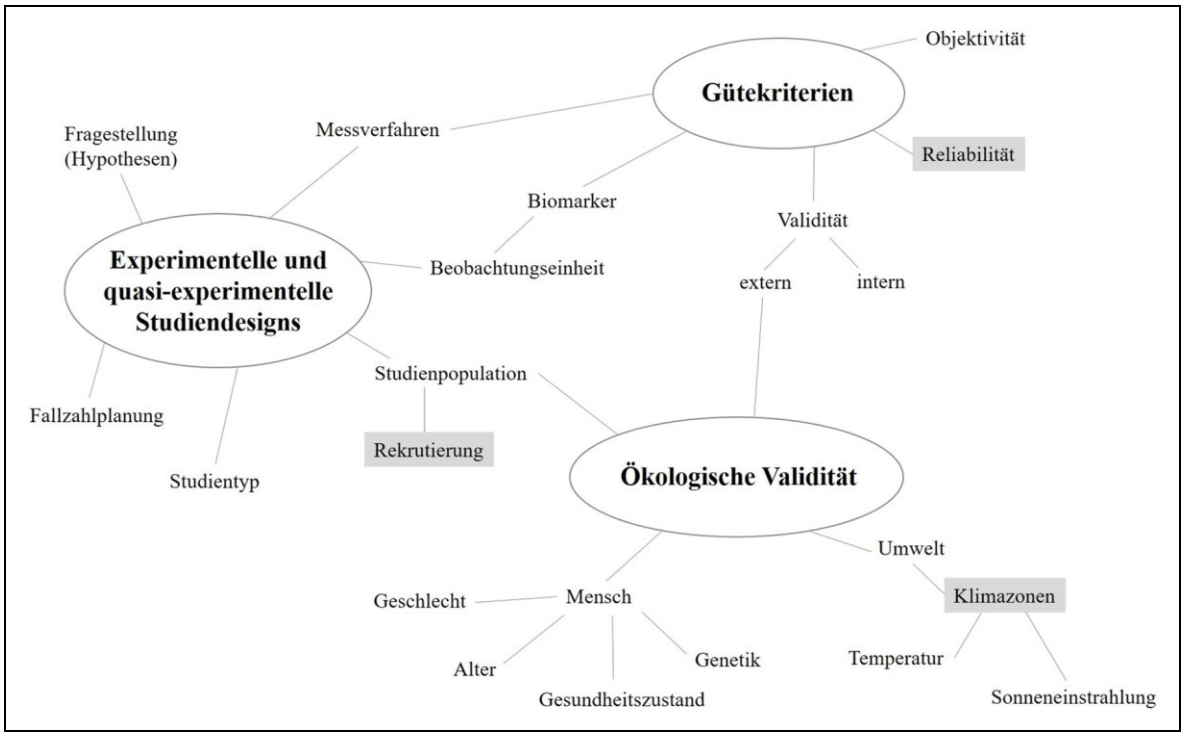


Abb. 2. Konzeptuelles Framework zum besseren Verständnis des Forschungsgegenstands: Methodenkritische Aspekte der Messung chronischer Schmerzen bei der Durchführung klinischer Studien

Als übergeordnete Zielsetzung setzt sich die vorliegende Dissertationsschrift kritisch mit methodischen Herausforderungen der Messung chronischer Schmerzen bei der Durchführung klinischer Studien auseinander. Auf Grundlage der in dieser Arbeit präsentierten Publikationen sollen folgende Zielsetzungen untersucht werden:

Zielsetzung I:

> *Analyse einer klinisch bedingten, signifikanten Abnahme der Schmerzintensität während der Rekrutierungsphase. Ziel soll es sein, dazu beitragende Faktoren zu untersuchen und Empfehlungen für Rekrutierungsverfahren für zukünftige klinische Studien zu chronischem Schmerz zu geben.*

Fragestellungen:

> *Welche Faktoren bedingen eine klinisch signifikante Schmerzreduktion während der Rekrutierung der Schmerzpatienten, noch vor Beginn der Intervention?*
>
> *Welche Empfehlungen für zukünftige klinische Studien resultieren daraus?*

Zielsetzung II:

Untersuchung der Test-Retest-Reliabilität sowie der Übereinstimmungsgrenzen des standardisierten QST-Protokolls über einen therapierelevanten Zeitraum von 10 Wochen bei gesunden Probanden.

Fragestellungen:

Wie reliabel sind die QST-Messungen im Zeitverlauf über 10 Wochen bei gesunden Probanden?

Welche Übereinstimmungsgrenzen können für das QST definiert werden?

Zielsetzung III:

Eine vergleichende Analyse des QST-Protokolls zwischen einer mitteleuropäischen und einer subarktischen Population zu klima-bedingten Unterschieden im Temperaturschmerzempfinden.

Fragestellung:

Zeigen sich klima-bedingte Unterschiede in den QST-Profilen einer subarktischen Population im Vergleich zu einer mitteleuropäischen Stichprobe?

2 Aktueller Forschungsstand

2.1 Schmerz und Nozizeption

Die International Association for the Study of Pain (IASP) definiert Schmerz folgendermaßen:
„Pain is an unpleasant sensory and emotional experience with actual or potential tissue damage or described in terms of such damage" (Schulte am Esch & Bause, 2011, S. 593). Demnach ist Schmerz ein unangenehmes Sinnes- und Gefühlserlebnis, das mit aktueller oder potentieller Gewebsschädigung verknüpft ist oder mit Begriffen einer solchen Schädigung beschrieben wird (Kröner-Herwig, 2004). Laut dieser Definition stellt Schmerz eine elementare Sinnesempfindung dar, die spezifisch beim Einwirken gewebeschädigender Reize ausgelöst wird. Diese Sinnesempfindung ist mit einem meist unangenehmen Gefühlserlebnis verbunden. Weiterhin kann der Definition entnommen werden, dass Schmerz auch als Ausdruck einer Gewebsschädigung empfunden wird, selbst wenn keine Gewebsschädigung vorliegt. Die Schmerzdefinition betont vor allem den emotionalen Aspekt als konstitutive Komponente des Schmerzgeschehens und unterscheidet damit Schmerz von anderen sensorischen Wahrnehmungsprozessen. Darüber hinaus hebt die Definition auch die Subjektivität des Schmerzerlebens und der Schmerzerfahrung hervor. Eine weitere Definition von Schmerz von McCaffery unterstreicht das individuelle und subjektive Empfinden von Schmerz. *„Schmerz ist, was immer ein Patient so bezeichnet und wann immer er dies tut"* (McCaffery et al., 1997, S. 12).
Im Gegensatz zur subjektiven Schmerzempfindung beschreibt der Begriff der Nozizeption die objektivierbaren peripheren und zentralen Vorgänge, die im Gehirn zur Schmerzwahrnehmung führen. Nozizeption ist die Aufnahme, Weiterleitung und Verarbeitung noxischer Reize durch das Nervensystem. Dadurch werden verschiedene Komponenten der Schmerzempfindung erzeugt, die von der Einwirkung eines noxischen Reizes auf den Körper bis hin zur bewussten Schmerzwahrnehmung und Bewertung eine Rolle spielen.

Abbildung 3 zeigt den Zusammenhang zwischen Nozizeption und Schmerz sowie die Interaktion der verschiedenen Komponenten der Schmerzempfindung. Die sensorisch-

diskriminative Komponente umfasst die Analyse der sensorischen Eigenschaften des noxischen Reizes. Informationen über Ort, Dauer, Art und Intensität des Reizes werden verarbeitet (Treede et al., 1999). Unter der affektiven Komponente wird die emotionale Bewertung der Schmerzen verstanden. Die sich daraus ergebende (meist negative) emotionale Reaktion äußert sich häufig in Form von Angst oder Unwohlsein (Weiss & Schaible, 2004). Die vegetative (autonome) Schmerzkomponente umfasst Reaktionen des vegetativen Nervensystems, die durch die Wahrnehmung von Schmerzen ausgelöst werden. Mögliche Reaktionen sind unter anderem der Anstieg von Blutdruck und Herzfrequenz, eine Erweiterung der Hautgefäße und der Pupillen sowie die Veränderung der Atmung (Birbaumer & Schmidt, 2010). Die durch den schmerzhaften Reiz ausgelösten Flucht- bzw. Schutzreflexe, durch die das betroffene Körperteil von der Schmerzquelle entfernt wird, werden unter der motorischen Schmerzkomponente verstanden. Dazu zählen beispielsweise Muskelverspannungen und Schonhaltungen (Weiss & Schaible, 2004; Birbaumer & Schmidt, 2010). Die Verarbeitung und Bewertung des Schmerzreizes bezeichnet die kognitive Komponente. Frühere Erfahrungen mit Schmerzen, die zu erwartenden Konsequenzen des Schmerzes sowie die wahrgenommene Verfügbarkeit von persönlichen oder externen Ressourcen, die die Möglichkeit zur Kontrolle des Schmerzes geben, können die Schmerzbewertung beeinflussen (Weisenberg, 1999; Weiss & Schaible, 2004). Kognitive Prozesse können Schmerzäußerungen, beispielsweise in Form von Mimik, auslösen (psychomotorische Komponente).

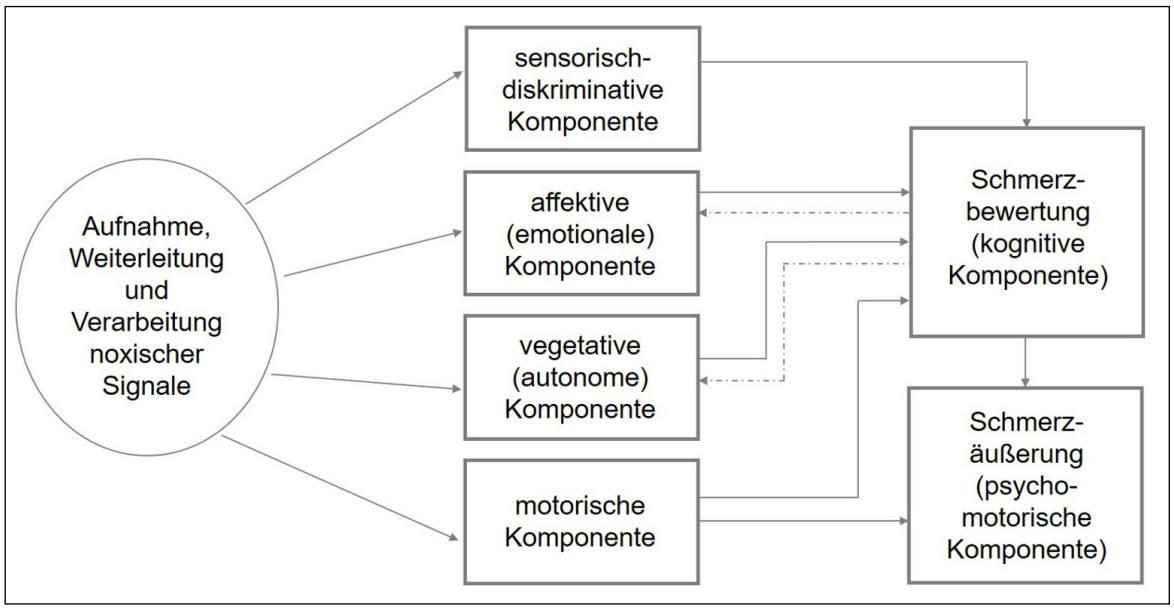


Abb. 3. Interaktion der verschiedenen Komponenten der Schmerzempfindung (mod. nach Schaible, 2010, S. 299)

2.1.1 Klassifikation von Schmerzen

Schmerzen lassen sich nach bestimmten Kriterien klassifizieren. Mutschler et al. (2007) sehen dafür folgende Einteilung vor:

- Art der Schmerzentstehung: physiologischer Nozizeptorschmerz
  pathophysiologischer Nozizeptorschmerz
  neuropathischer Schmerz
- Ort der Entstehung: somatischer Oberflächenschmerz
  somatischer Tiefenschmerz
  viszeraler Tiefenschmerz
- Dauer des Schmerzes: akuter Schmerz
  chronischer Schmerz

Nach Art der Schmerzentstehung kann der Schmerz in physiologischen und pathophysiologischen Nozizeptorschmerz und neuropathischen Schmerz differenziert werden. Die Ursache des physiologischen Nozizeptorschmerzes liegt in der Einwirkung gewebeschädigender Reize auf normales Gewebe und der damit ausgelösten Schmerzwahrnehmung. Seiner Funktion entsprechend stellt er eine Warnung vor Gewebeschädigungen dar, so dass unwillkürlich Gegenmaßnahmen eingeleitet werden können wie beispielsweise das rasche Wegziehen der Hand, wenn man versehentlich auf die heiße Herdplatte gefasst hat. Als Folge pathologischer Organveränderungen (z. B. Entzündungen) wird der pathophysiologische Nozizeptorschmerz ausgelöst (Schaible, 2010). Geht der Schmerz mit einer Schädigung im Bereich peripherer Nerven oder des zentralen Nervensystems einher, spricht man von neuropathischem Schmerz. Mögliche Ursachen stellen beispielsweise Nervenschäden durch Lagerungsfehler, Amputationen, Tumorinfiltrationen, Entzündungen und Schädel-Hirn-Traumata dar. Neuropathischer Schmerz kann mit einem brennenden, stechenden bzw. bohrenden Charakter wahrgenommen werden (Bone et al., 2010).

Darüber hinaus lassen sich Schmerzen aufgrund ihres Entstehungsorts in somatischen (Oberflächenschmerz und Tiefenschmerz) und viszeralen Schmerz unterscheiden. Die

noxische Reizung der Haut löst einen somatischen Oberflächenschmerz aus. Dieser sogenannte erste Schmerz wird als hell und gut lokalisierbar empfunden und klingt nach Beendigung des Reizes schnell ab. Der somatische Oberflächenschmerz besitzt die Aufgabe, Schutz- und Fluchtreflexe auszulösen (Schaible, 2010). Anschließend folgt ein zweiter Schmerz mit dumpfem oder brennendem Charakter, der nur langsam abklingt. Der ebenfalls als dumpf wahrgenommene, von Muskeln, Gelenken, Knochen und vom Bindegewebe kommende Schmerz wird als somatischer Tiefenschmerz definiert. Dieser lässt sich nur schwer lokalisieren und neigt dazu, in die Umgebung auszustrahlen (Mutschler et al., 2007). Somatische Schmerzen werden als dumpf, bohrend, ziehend oder stechend beschrieben und oft durch Bewegung oder Druck verstärkt (Grond & Radbruch, 2002). Der viszerale Schmerz entsteht bei Erkrankungen innerer Organe. Er ist schlecht lokalisierbar und strahlt oft in die Umgebung aus (Huppelsberg & Walter, 2009). Viszerale Schmerzen werden vorwiegend als krampfartig oder kolikartig bezeichnet (Grond & Radbruch, 2002).

Letztlich können Schmerzen nach Dauer des Auftretens in akuten und chronischen Schmerz klassifiziert werden. Der akute Schmerz ist durch seine begrenzte zeitliche Wahrnehmungsdauer charakterisiert und hält so lange an, wie die ursächliche Störung oder deren Folgen bestehen. Meistens sind akute Schmerzen auf eine Dauer von wenigen Stunden oder Tagen beschränkt (Grond & Radbruch, 2002). Die Bedeutung des akuten Schmerzes liegt in der Warn- und Schutzfunktion, um die Unversehrtheit des Körpers zu sichern (Bone et al., 2010). Schulte am Esch und Bause (2011) sprechen von einem Frühwarnsystem des Körpers, welches unter anderem ein schmerzvermeidendes Verhalten fördert. Chronische Schmerzen haben ihre Warn- und Schutzfunktion verloren und können sich zu einem eigenständigen Krankheitsbild entwickeln. Darüber hinaus wird in der Literatur häufig von chronischen Schmerzen gesprochen, wenn der Schmerz länger als drei oder sechs Monate anhält. Die Definition chronischer Schmerzen ausschließlich auf die zeitliche Komponente zu beschränken, wird der Komplexität und der Individualität des Chronifizierungsprozesses jedoch nicht ausreichend gerecht. Chronische Schmerzen sind häufig multikausal, das heißt sie haben meist mehrere schmerzauslösende Ursachen (Schulte am Esch & Bause, 2011). Langandauernde Schmerzen können für den Betroffenen zu einer starken emotionalen, psychischen und sozialen Belastung werden (Bone et al., 2010).

2.1.2 Epidemiologische Aspekte chronischer Schmerzen

Chronische Schmerzen stellen weltweit eine große Herausforderung für die Gesundheitssysteme dar. Die Prävalenzraten chronischer Schmerzen zeigen eine große Spannweite auf. 2012 berichtete die Deutsche Schmerzgesellschaft, dass 14 Mio. Deutsche (17 %) an nichttumorbedingten chronischen Schmerzen leiden. In einer weiteren Studie aus dem Jahr 2012/2013 klagten 32,9 % der Teilnehmer über chronische Schmerzen, 5,4 % gaben einen chronischen Schmerz mit assoziierten körperlichen und sozialen Beeinträchtigungen und 2,3 % einen chronischen Schmerz mit assoziierten körperlichen, seelischen und sozialen Beeinträchtigungen an (Häuser et al., 2013). Laut den Ergebnissen der bislang größten epidemiologischen Untersuchung zu chronischem Schmerz in Europa litten im Jahr 2005 ca. 19 % der erwachsenen Bevölkerung an chronischen Schmerzen, 34 % davon unter starken, 46 % unter dauerhaften Schmerzen. 19 % der Schmerzpatienten hatten ihren Arbeitsplatz aufgrund der Einschränkung verloren, 13 % mussten den Beruf wechseln. Bei einem Fünftel der Betroffenen wurde darüber hinaus die Diagnose einer Depression gestellt (Breivik et al., 2006).

Gemäß eines Reviews der Global Burden of Disease stellen chronische Schmerzen einen der Hauptgründe für krankheitsbedingte Einschränkungen, Behinderungen und generelle Minderungen der Lebensqualität dar. Eine ältere, europäische Schmerzbefragungsstudie ergab, dass Personen mit chronischen Schmerzen meistens länger als 7 Jahre daran leiden. Darüber hinaus berichteten ein Sechstel der Befragten, dass sie manchmal aufgrund ihrer Schmerzen sterben möchten. 40 % der Befragten wiesen darauf hin, dass sie durch ihre Schmerzen stark im Lebensalltag eingeschränkt sind. 27 % haben aufgrund ihrer chronischen Schmerzen Beziehungsprobleme (Breivik et al., 2006). Chronische Schmerzen gehen mit einer relevanten Einschränkung der Lebensqualität einher. Dabei korreliert die Schmerzintensität stark negativ mit der Lebenszufriedenheit (Bellach et al., 2000). Die durch chronische Schmerzen entstehenden Gesundheitskosten sind enorm und stellen weltweit ein bedeutendes gesundheitsökonomisches Problem dar.

2.2 Diagnostik chronischer Schmerzen

> *„Ziel von Schmerzdiagnostik ist es, die komplexen Wechselwirkungen biologischer, psychologischer und sozialer Einflüsse bei der Entstehung, Aufrechterhaltung und Remission zu verstehen.“* (Nilges & Rief, 2016, S. 35)

Unter anderem zählen schmerzbezogene Anamnesen, Schmerzfragebögen, Messungen der Schmerzstärke, Schmerztagebücher sowie Messungen der Schmerzempfindlichkeit zur Schmerzdiagnostik. So lassen sich beispielsweise die mit Schmerzen einhergehenden Symptome mittels standardisierter Fragebögen erfassen. Dabei können Zeichnungen des menschlichen Körpers die betroffene(n) Körperstelle(n) und die Ausbreitung der Schmerzen veranschaulichen. In der klinischen Praxis kommen häufig der „Deutsche Schmerzfragebogen“ sowie der „painDetect-Fragebogen“ zum Einsatz. In einem Schmerztagebuch können täglich Informationen zum Auftreten (Häufigkeit) sowie zur Stärke (Intensität) und Dauer der Schmerzen gesammelt werden, um einen Überblick über den Symptom- und Behandlungsverlauf zu erhalten (Radbruch & Grond, 2002).

2.2.1 Messung der Schmerzstärke

Die Schmerzmessung im Sinne der Objektivierung der Schmerzstärke steht insofern vor einer besonderen Herausforderung, als dass nur das Schmerzerleben durch die Angabe des Patienten objektivierbar ist, nicht aber die reine Nozizeption. Das Erleben von Schmerzen ist individuell sehr unterschiedlich und wird unter anderem durch die Schmerzerfahrung und die psychische Verfassung des Patienten beeinflusst. Zur Messung des Schmerzes kommen in der klinischen Praxis unter anderem eindimensionale Schmerzskalen zum Einsatz. McCaffery (1997) formuliert die Intention von Schmerzskalen wie folgt:

> *„die Beschreibung des Patienten über das Ausmaß der Schmerzen in Zahlen oder Worte zu übertragen, welche eine möglichst objektive Beschreibung für eine subjektive Erfahrung liefern soll.“* (McCaffery et al., 1997, S. 15)

Folgende Skalen werden zur Messung der Schmerzstärke eingesetzt: visuelle Analogskala (VAS), verbale Ratingskala (VRS) sowie die numerische Ratingskala (NRS). Je nach gewählter Schmerzskala gibt der Patient eine (Selbst-) Einschätzung seiner (aktuellen) Schmerzintensität mittels Deskriptor oder Zahl an (Abbildung 4).

Mithilfe der NRS ordnen Patienten ihrer empfundenen Schmerzstärke eine Zahl zu. In der klinischen Praxis kommen häufig elfstufige Skalen (Einteilung von 0 bis 10) bzw. Skalen mit einer Einteilung von 0 bis 100 zum Einsatz, wobei 0 jeweils „keine Schmerzen“ und die höchste Zahl „unerträgliche Schmerzen“ oder „stärkster vorstellbarer Schmerz“ bedeutet. Unter Anwendung der VAS markiert der Patient seine subjektive Schmerzempfindung auf einer 100 mm langen Linie mit den Endpunkten „kein Schmerz“ und „stärksten vorstellbaren

Schmerzen“ (Bone et al., 2010). Die VRS stellt eine verbal deskriptive Schmerzskala dar. Der Patient drückt seine empfundene Schmerzstärke durch Worte (Deskriptoren) aus. Typischerweise werden fünfteilige VRS verwendet, wobei der Schmerz von „kein Schmerz“, „mäßig“, „mittelstark“, „stark“ bis „stärkster vorstellbarer Schmerz“ bewertet werden kann (Schulte am Esch & Bause, 2011). Aufgrund der Unterteilung in fünf Schmerzstärken ist die VRS weniger sensibel als die NRS und die VAS (Downie et al., 1978).

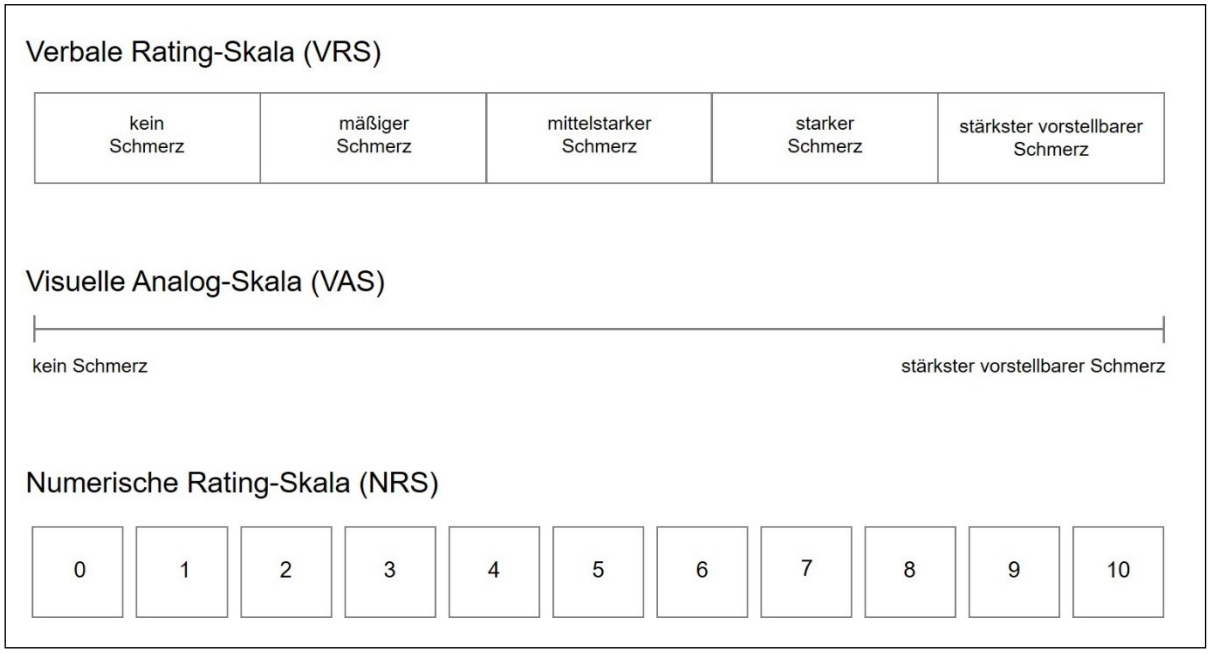


Abb. 4. Skalen zur Objektivierung der Schmerzintensität (Kompendium der medikamentösen Schmerztherapie, Beubler, 2016, S. 7)

2.2.2 Messung der Schmerzempfindlichkeit mittels der Quantitativen Sensorischen Testung (QST)

Für die Messung der Schmerzempfindlichkeit dient die Erhebung verschiedener Wahrnehmungs- und Schmerzschwellen. Im Rahmen eines vom Bundesministerium für Bildung und Forschung geförderten Projektes wurde dazu ein standardisiertes Testprotokoll vom Deutschen Forschungsverbund Neuropathischer Schmerz (DFNS) entwickelt: die Quantitativ Sensorische Testung (QST).

QST ist ein psychophysisches Testverfahren des somatosensorischen Nervensystems. Die vollständige Testbatterie der QST nach DFNS Standardprotokoll umfasst die Durchführung

von 7 verschiedenen Subtests zur Erhebung von 13 somatosensorischen Testparametern, die in Tabelle 1 aufgelistet sind.

Tab. 1. *Übersicht der 13 somatosensorischen QST-Parameter.*

thermisch	Kältedetektionsschwelle	(CDT)
	Wärmedetektionsschwelle	(WDT)
	thermische Unterschiedsschwelle	(TSL)
	paradoxe Hitzeempfindungen	(PHS)
	Kälteschmerzschwelle	(CPT)
	Hitzeschmerzschwelle	(HPT)
mechanisch	mechanische Detektionsschwelle	(MDT)
	mechanische Schmerzschwelle	(MPT)
	mechanische Schmerzsensitivität	(MPS)
	dynamischen mechanischen Allodynie	(DMA)
	Wind-up-Phänomen	(WUR)
	Vibrationsschwelle	(VDT)
	Druckschmerzschwelle	(PPT)

Abbildung 5 zeigt den schematischen Ablauf des standardisierten QST-Protokolls. Die exakte Methodik der QST nach standardisierten DFNS Protokoll wurde bereits vielfach publiziert (Rolke et al., 2006a und 2006b; Krumova et al., 2012) und ist ebenso Bestandteil der in dieser Dissertation enthaltenen Publikation II (Nothnagel et al., 2017). Für eine ausführliche Beschreibung der QST-Methodik wird daher auf die Publikation II verwiesen.

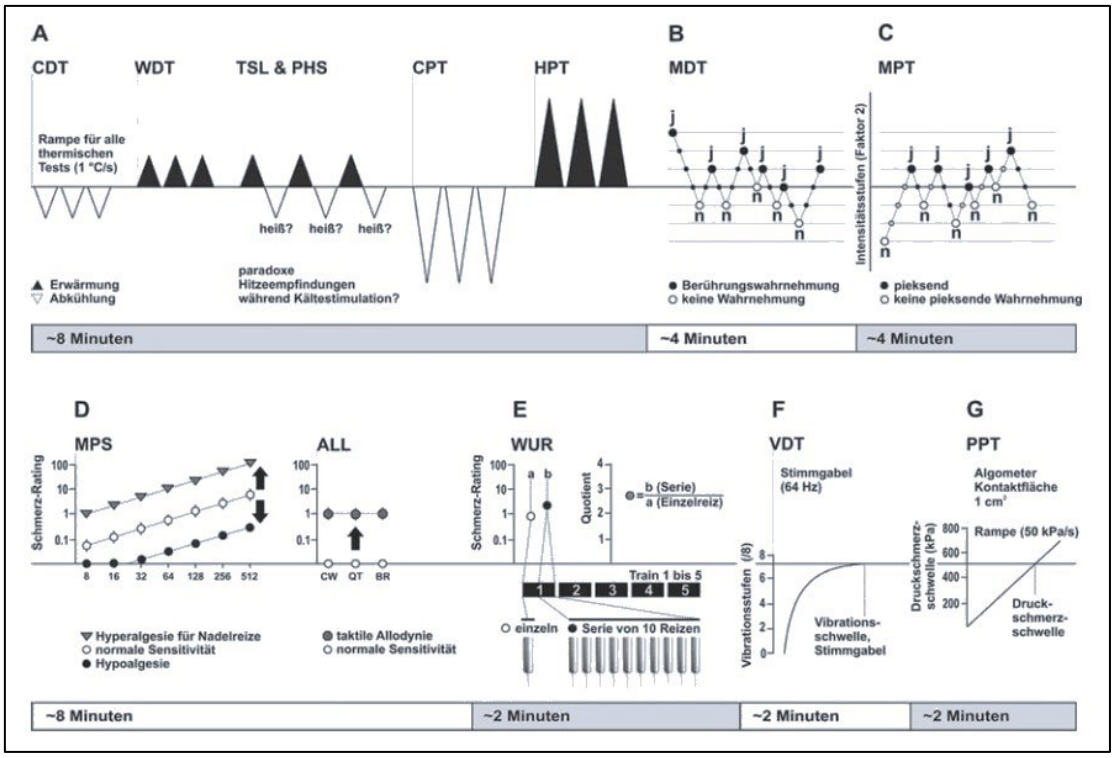


Abb. 5. Standardisiertes QST-Protokoll. Methodik der Testbatterie mit 7 Subtests zur Bewertung von 13 sensorischen Parametern (Rolke et al., 2006b, S. 79)

Durch kalibrierte thermische und mechanische Reize erlaubt die QST die Erfassung des Funktionszustands aller primär-afferenten Fasersysteme (Baron & Tölle, 2009). Konventionelle neurophysiologische Messverfahren (z. B. Nervenleitgeschwindigkeit-Messung) untersuchen ausschließlich die Funktion der großen sensiblen, dick myelinisierten Nervenfasern (Aβ-Fasern), erlauben aber keine Aussagen über dünn myelinisierte und unmyelinisierte Nervenfasern (Aδ- und C-Fasern). Die QST kann darüber hinaus sowohl die Aβ-Fasern als auch die Aδ- und C-Fasern ansprechen. Die Testung der Schmerzschwellen und der thermischen Schwellen bildet die Funktionen der dünnen Nervenfasern und des spinothalamischen Trakts im Vorderseitenstrang (Aδ- und C-Fasern) ab. Mögliche Funktionsstörungen der dick myelinisierten Aβ-Fasern und des Hinterstrangsystems werden durch die Erhebung der mechanischen Detektionsschwellen geprüft. Tabelle 2 zeigt eine Übersicht der QST-Subtests zur Abbildung der Nervenfasern sowie der klinischen Zeichen. Durch die QST können Sensibilitätsveränderungen im Sinne von Funktionssteigerung und Funktionsverlust detektiert werden. Beim Vorliegen einer gesteigerten Empfindlichkeit (Hyperalgesie, Hyperpathie und Allodynie) spricht man von sensorischen Positivzeichen, beim Vorliegen eines Sensibilitätsausfalls (Hypästhesie und Hypalgesie) von sensorischen

Negativzeichen. Diese klinischen Zeichen können auf den vom Schmerz ausgehenden neurobiologischen Mechanismus hinweisen (periphere und zentrale Sensibilisierung sowie Deafferenzierung).

Tab. 2. *Übersicht der somatosensorischen Modalitäten, QST-Subtests, periphere Bahnen und klinische Zeichen (mod. nach Hansson et al., 2007; Backonja et al., 2013; Spohn et al., 2013).*

Modalität	QST-Subtest	Afferenz	positive Zeichen	negative Zeichen
<u>thermisch</u>				
Kälte	CDT, WDT, TSL, PHS	Aδ- Fasern	thermische Hyperästhesie	thermische Hypästhesie
Wärme		C-Fasern		
Kälteschmerz	CPT	Aδ, C-Fasern		
Hitzeschmerz	HPT	Aδ, C-Fasern	thermische Hyperalgesie	thermische Hypalgesie
<u>mechanisch</u>				
Berührung	MDT	Aβ-Fasern		mechanische Hypästhesie
Nadelreize	MPT, MPS, WUR	Aδ, C-Fasern	mechanische Hyperalgesie	mechanische Hypalgesie
nicht-nozizeptive Reize	DMA	Aβ-Fasern	Allodynie	
Vibration	VDT	Aβ-Fasern		mechanische Hypästhesie
stumpfer Druck	PPT	Aδ, C-Fasern	mechanische Hyperalgesie	mechanische Hypalgesie

Die DFNS führt eine ständig wachsende Referenzdatenbank mit alters- und geschlechtsadjustierten Normwerten für Gesicht, Hand und Fuß (Rolke et al., 2006a; Magerl et al., 2010). Weitere Referenzwerte stehen für den Rumpf (Pfau et al., 2014) sowie für krankheitsspezifische somatosensorische Profile (Blankenburg et al., 2010; Magerl et al., 2010; Maier et al., 2010) zur Verfügung.

QST ist auf die Kooperation bzw. Mitarbeit des Patienten angewiesen. Mangelnde Aufmerksamkeit, Motivation, kognitive Beeinträchtigung, psychiatrischer Komorbiditäten

oder Vortäuschungen können die Ergebnisse beeinflussen (Hansson et al., 2007; Backonja et al., 2013).

In den letzten Jahrzehnten hat die Methode der QST zunehmende Aufmerksamkeit erfahren, sowohl in der klinischen Diagnostik als auch zu wissenschaftlichen Forschungszwecken. Insbesondere auf dem Gebiet der neuropathischen Schmerzen hat sich die Methode der QST zu einem etablierten Analyseverfahren in der Schmerzdiagnostik entwickelt. In weltweiten Schmerzstudien wird der vom DFNS entwickelte Testalgorithmus als Goldstandard verstanden und angewandt.

QST kommt vor allem als diagnostisches Instrument bei Rückenschmerzen (Freynhagen, 2008; Mücke et al., 2014), beim komplexem regionalem Schmerzsyndrom (CRPS) (Huge et al., 2008; Eberle et al., 2009; Mücke et al., 2014), bei Small-fiber-Neuropathien (Maag et al., 2008; Scherens et al., 2009) und bei Diabetes zum Einsatz (Pavlaković & Petzke, 2010; Birklein & Sommer, 2013; Uddin & MacDermid, 2016). Die QST allein erlaubt zwar keine Schmerzdiagnostik, liefert aber wichtige Zusatzinformationen zum individuellen Schmerzprofil des Patienten. Daraus kann beispielsweise auf eine verminderte Nervenfunktion durch eine Nervenschädigung geschlossen werden. Alternativ ergeben sich Hinweise auf eine Nervenüberempfindlichkeit bei anderen Schmerzerkrankungen, die ohne eine bedeutsame Nervenschädigung entstehen (Kopfschmerz, muskulärer Rückenschmerz, Fibromyalgie) (Mücke et al., 2014). In der klinischen Forschung ist die QST ein wichtiges Instrument zur Charakterisierung definierter Patientenkollektive und zur Evaluation von Interventionen.

2.3 Interventionsstudien zum chronischen Schmerz: methodenkritische Aspekte

2.3.1 Schmerzstärke als Einschlusskriterium in klinischen Studien

Die Einschätzung der Schmerzintensität mittels Schmerzskala wird häufig zur Verlaufs- und Therapiekontrolle in der klinischen Praxis sowie in wissenschaftlichen Studien genutzt. Darüber hinaus wird die empfundene Schmerzstärke häufig auch als ein Einschluss- bzw. Ausschlusskriterium für Interventionsstudien mit Schmerzpatienten definiert. So werden beispielsweise in bestimmten Interventionsstudien nur Schmerzpatienten mit einer durchschnittlichen Schmerzintensität von > 40 auf einer NRS bzw. VAS von 0 bis 100

eingeschlossen. Der Ein- bzw. Ausschluss potentieller Schmerzpatienten ist Teil der Rekrutierungsphase. Je nach Art der klinischen Studie kann dieser Rekrutierungsprozess sehr zeitaufwendig sein. Dass die Rekrutierung in klinischen Studien von typischen Problemen begleitet wird, wurde hinlänglich publiziert (Aitken et al., 2003; Gul & Ali, 2010; Probstfield & Frye, 2011). Darüber hinaus ist aber besonders bei Patienten mit einer schwankenden Schmerzsymptomatik und somit auch mit einer schwankenden Schmerzintensität der Zeitpunkt des Einschlusses in die Studie entscheidend. Je nach Erkrankung entstehen individuell unterschiedliche Krankheitsverläufe (natural course of disease). Bei akuten Erkrankungen, beispielsweise einem Herzinfarkt, zeigt sich ein aufsteigender Verlauf der Symptomatik. Dagegen weisen chronische Erkrankungen häufig Schwankungen im Krankheitsverlauf auf, wie diese typischerweise bei Migräne, chronischen Nacken- und Rückenschmerzen oder rheumatischen Erkrankungen vorkommen. Diese Veränderungen hängen stark von den unterschiedlichen Stadien einer Erkrankung ab und bestimmen die individuelle Krankheitskurve. Für Studien mit chronischen Schmerzpatienten stellt dieser schwankende Krankheitsverlauf deshalb eine besondere Herausforderung dar.

Darüber hinaus können sich auch das Phänomen der Regression zum Mittelwert (Regression-to-the-mean) und der Hawthorne-Effekt auf die Daten und Ergebnisse auswirken. So lassen sich in klinischen Studien möglicherweise Änderungen, bezogen auf das Verhalten der Daten, auf das Phänomen der Regression zum Mittelwert und Änderungen, bezogen auf das Verhalten der Patienten, auf den Hawthorne-Effekt zurückführen (Kleist, 2006). Der Hawthorne-Effekt beschreibt, dass Teilnehmer einer Studie ihr natürliches Verhalten ändern, weil sie wissen, dass sie an einer Studie teilnehmen und unter Beobachtung stehen werden. In klinischen Studien kann dieser Effekt beispielsweise zu einer Überschätzung der Wirksamkeit in einer Kontrollgruppe führen und den Nachweis der Wirksamkeit einer effektiven Therapie erschweren bzw. sogar verhindern (Kleist, 2006). Das Phänomen der Regression zum Mittelwert rückt dann in den Fokus, wenn Messungen grundsätzlich keine konstanten Werte ergeben, sondern eine Variabilität aufweisen, beispielsweise aufgrund intraindividueller Schwankungen, situativer Anpassungen oder aufgrund von Messungenauigkeiten. In klinischen Studien erfolgt die Auswahl der Studienpopulation häufig auf Grundlage extremer Ausgangswerte, entsprechend sehr hoher oder sehr niedriger Baseline-Werte. Diese unterliegen keiner Normalverteilung, sondern stellen das extreme Ende der Variabilität in der untersuchten

Patientenpopulation dar. Die in einer Wiederholungsmessung erhobenen Werte werden mit einer hohen Wahrscheinlichkeit näher am eigentlich individuellen bzw. am Gruppenmittelwert liegen. Damit lässt das Regression zum Mittelwert-Phänomen die natürliche Variabilität von Wiederholungsmessungen wie eine tatsächliche Verbesserung der Messdaten aussehen.

Eine umfangreiche Erörterung der Phänomene Natural course of disease, Regression-to-the-mean und des Hawthorne-Effekts sowie ihr Zusammenwirken in klinischen Studien befinden sich in der Publikation I.

2.3.2 Chronischer Schmerz im zeitlichen Verlauf: Wie reproduzierbar ist die Messung der Schmerzempfindlichkeit?

Die Analyse der Schmerzempfindlichkeit mittels des standardisierten QST-Protokolls hat sich in den letzten 10 Jahren zu einem Goldstandard in der Schmerzforschung und -diagnostik entwickelt. Aufgrund der Unterscheidungseigenschaften der QST kann davon ausgegangen werden, dass diese Methode ebenfalls in der Lage ist, Veränderungen über die Zeit zu erfassen. Dies ist relevant, um sowohl pathologische als auch funktionelle Veränderungen identifizieren zu können und könnte möglicherweise auch für die Quantifizierung von Behandlungseffekten in klinischen Studien zu chronischen Schmerzen von Bedeutung sein. Eine entscheidende Voraussetzung für die Nützlichkeit der QST als Outcome-Parameter in klinischen Studien ist jedoch, dass wiederholte Messungen der QST über die Zeit stabil sind. Vor allem bei chronischen Erkrankungen mit einem schwankenden Krankheitsverlauf sind Zuverlässigkeitsanalysen über einen längeren Zeitraum notwendig. Um dies postulieren zu können, ist eine Test-Retest-Reliabilitätsanalyse der QST-Parameter über einen längeren, therapierelevanten Zeitraum essenziell. Zur Analyse und Beurteilung der Test-Retest-Zuverlässigkeit dient die Berechnung des Intraclass-Korrelationskoeffizienten (ICC) (Werner et al., 2013). Die Analyse des ICC basiert dabei auf der Berücksichtigung der durchschnittlichen Differenzen der Messwerte, der sogenannten Verzerrung (Bias) (Grouven et al., 2007). Dies ist zur Beurteilung der Übereinstimmung zweier Messungen nicht ausreichend (Grouven et al., 2007). Bland und Altman (2012, pp. 183) formulieren dieses methodische Problem bei wiederholten Messungen wie folgt: *„If one measurement is always twice as big as the other, they are highly correlated, but they do not agree”.* So kann der ICC auch bei einer hohen

Übereinstimmung der wiederholten Messungen kleine Werte annehmen, wenn die Streuung zwischen den wiederholten Messungen im Verhältnis zur Streuung der Daten innerhalb einer Messung klein ist (Grouven et al., 2007). Von entscheidender Bedeutung ist also die Betrachtung der Streuung der Differenzen der einzelnen Messwertpaare. Dafür stellt der Bland Altmann Plot ein geeignetes, graphisch-statistisches Verfahren dar, dass die Verzerrung und die Streuung der Daten berücksichtigt. Dabei werden die Differenzen der Messwertpaare gegen die Mittelwerte der Messwertpaare aufgetragen. Zusätzlich wird der Mittelwert aller Differenzen als horizontale Linie übertragen. Darüber hinaus werden der Mittelwert aller Differenzen ± 1,96 x Standardabweichung der Differenzen als Linien in das Diagramm eingezeichnet. Diese Linien definieren die sogenannten Übereinstimmungsgrenzen (Limits of Agreement, LoA). Im Plot wird sichtbar, inwiefern beide Messungen schwanken, ob die eine Messung systematisch höhere Werte oder niedrigere Werte als die andere Messung misst (systematischer Messfehler). Des Weiteren lassen sich eventuelle Abhängigkeiten erkennen, das heißt ob eine größere Schwankungsbreite beispielsweise von der Höhe der Werte abhängt (Grouven et al., 2007).

Eine umfangreiche Analyse der Test-Retest-Reliabilität sowie der Übereinstimmung zweier QST-Messungen wird in der Publikation II dieser Dissertationsschrift diskutiert.

2.3.3 Fundamentale Einflüsse der Schmerzwahrnehmung

Experimentelle Untersuchungen zum chronischen Schmerz belegen, dass Frauen schmerzempfindlicher sind als Männer. Frauen leiden unter intensiveren und länger andauernden Schmerzen. Zudem zeigen epidemiologische Studien eine höhere Prävalenz chronischer Schmerzen bei Frauen (Dieleman et al., 2008; Icenhour et al., 2015). Weitere Forschungsarbeiten zeigen eine veränderte Schmerzwahrnehmung abhängig vom Alter der Patienten sowie von der Schmerzlokalisation (Meh & Denišlič, 1994). Bezogen auf die Messung der Schmerzempfindlichkeit mittels QST werden die Faktoren Geschlecht, Alter und Schmerzlokalisation durch die DFNS Referenzdaten, die sowohl alters- und geschlechtsnormiert sind, als auch je nach Schmerzregion unterschieden werden, berücksichtigt (Magerl et al., 2010; Pfau et al., 2014).

Darüber hinaus sind bereits Einflüsse der ethnischen Herkunft auf die Schmerzwahrnehmung und Schmerzschwellen bekannt (Bates et al., 1993; Edwards &

Fillingim, 1999; Edwards et al., 2001; Dawson & List, 2009). Auch das Geschlecht und der berufliche Status des Untersuchers scheinen in den Schmerzberichten der Teilnehmer einen Einfluss zu haben (Levine & de Simone, 1991; Kállai et al., 2004; Aslaksen et al., 2007). Im Rahmen der Durchführung klinischer Studien zum chronischen Schmerz sind unter anderem diese fundamentalen Faktoren zum Schmerz zu berücksichtigen.

Nach bisherigem Kenntnisstand fehlen bislang Untersuchungen zu Einflüssen von Umwelt und Klima auf die Schmerzwahrnehmung. So gibt es keine Untersuchungen dazu, inwieweit beispielsweise höhere Jahresdurchschnittstemperaturen, wie sie in mitteleuropäischen Regionen vorherrschen, Einfluss auf die thermische Schmerzempfindung nehmen. Des Weiteren fehlen Studien zum möglichen Einfluss von dauerhaft niedrigeren Temperaturen, die beispielsweise in subarktischen Regionen vorherrschen. Werden Schmerzen, vor allem thermische Schmerzreize, in subarktischen Regionen anders toleriert als in mitteleuropäischen Regionen? Diese Fragestellung wird in Publikation III dieser Arbeit diskutiert.

3 Eigene publizierte Originalarbeiten

Mithilfe der nachfolgenden Publikationen werden die unter Punkt 1 genannten Zielstellungen analysiert und die entsprechenden Fragestellungen beantwortet. Die abschließende Diskussion dieser Arbeit erörtert die Ergebnisse der Publikationen kritisch und zeigt die Implikationen der Ergebnisse für die Forschung und die klinische Praxis. Des Weiteren werden Anregungen für die weiterführende Forschung klinischer Studien zum chronischen Schmerz gegeben.

3.1 Publikation I: Recruitment and inclusion procedures as "pain killers" in clinical trials?

Titel: Recruitment and inclusion procedures as „pain killers" in clinical trials?

Autoren: Helen Nothnagel, Martha Brown Menard, Gunnvald Kvarstein, Arne Johan Norheim, Thomas Weiss, Christian Puta, Scott Mist, Frauke Musial

Journal: *Journal of Pain Research (2019); 12:* 2027-2037. DOI: 10.2147/JPR.S204259. IF 2.386. Veröffentlichtes Manuskript

Die Rekrutierung und der Einschluss von Probanden sind zeitkritische Faktoren in klinischen Studien. Dies gilt insbesondere für Studien, die Patienten mit fluktuierenden Symptommustern untersuchen wie beispielsweise Patienten mit chronischen Nackenschmerzen. In einer Machbarkeitsstudie zu Nackenschmerzen wurde eine klinisch relevante Abnahme der Schmerzwerte innerhalb des Rekrutierungszeitraums herausgefunden. Dieser Beitrag analysiert das Phänomen und gibt Empfehlungen für Rekrutierungsverfahren in klinischen Studien zu Schmerzen. Dazu wurden die Veränderungen der Schmerzintensitäten von 44 chronischen Nackenschmerzpatienten (6 Männer und 36 Frauen; mittleres Alter: 45,3 ± 13,2 Jahre) zwischen dem ersten telefonischen Kontakt und der Baseline-Untersuchung analysiert. Einschlusskriterium war eine mittlere Schmerzintensität von > 40 auf einer NRS von 0 bis 100 während der letzten drei Monate. Statistische Analysen wurden mit ANOVA und parametrischen/nicht-

parametrischen Korrelationskoeffizienten durchgeführt. Der durchschnittliche Wert der Schmerzintensität sank signifikant von 60,3 ± 13,3 beim ersten telefonischen Kontakt auf 38,1 ± 21,7 bei der Baseline-Untersuchung. Dies entspricht einer relativen Veränderung von 36,8 %. Es wurde eine schwache, aber signifikante negative Korrelation zwischen der Anzahl der Tage zwischen den Beurteilungen und den Änderungen in der Schmerzbewertung gefunden. Darüber hinaus zeigte sich eine positive Korrelation zwischen der Veränderung der Schmerzintensität und dem Schmerzniveau beim ersten Kontakt, was darauf hindeutet, dass die im Laufe der Zeit abnehmenden Schmerzbewertungen auch von der anfänglichen Schmerzbewertung abhängig waren. Die klinisch signifikanten Veränderungen der Schmerzintensität waren schwach von der Wartezeit und mäßig von der anfänglichen Schmerzintensität abhängig, was auf eine Regression zum Mittelwert hindeutet. Der natürliche Verlauf der Erkrankung und der Hawthorne-Effekt werden ebenfalls als beitragende Faktoren diskutiert.

3.2 Publikation II: How stable are quantitative sensory testing measurements over time? Report on 10-week reliability and agreement of results in healthy volunteers.

Titel: How stable are quantitative sensory testing measurements over time? Report on 10-week reliability and agreement of results in healthy volunteers.

Autoren: Helen Nothnagel, Christian Puta, Thomas Lehmann, Philipp Baumbach, Martha Brown Menard, Brunhild Gabriel, Holger Gabriel, Thomas Weiss, Frauke Musial

Journal: *Journal of Pain Research (2017); 10*: 2067-2078. DOI: 10.2147/JPR.S137391. IF 2.386. Veröffentlichtes Manuskript

Die QST ist ein diagnostisches Hilfsmittel zur Beurteilung des somatosensorischen Systems. Um QST als Ergebnismaß für klinische Studien etablieren zu können, ist die Frage entscheidend, wie ähnlich die Messungen über die Zeit sind. Daher werden die Zuverlässigkeit und Übereinstimmung des standardisierten QST-Protokolls des DFNS über einen therapierelevanten Zeitraum von 10 Wochen getestet. QST wurde dafür bei

22 gesunden Probanden (10 Männer und 12 Frauen; mittleres Alter: 46,6 ± 13,0 Jahre) zweimal auf dem unteren Rücken und dem Handrücken der dominanten Hand durchgeführt, wobei die Messungen im Abstand von 10,0 ± 2,9 Wochen stattfanden. Alle Messungen wurden von einem Untersucher durchgeführt. Zur Untersuchung der Langzeitzuverlässigkeit und -übereinstimmung der QST wurden die Unterschiede zwischen den beiden Messungen, die Korrelationskoeffizienten, die Intraclass-Korrelationskoeffizienten (ICCs), die Bland-Altman-Plots (Grenzen der Übereinstimmung) und der Standardfehler der Messung verwendet. Die meisten Parameter der QST waren über 10 Wochen bei gesunden Probanden zuverlässig: Nahezu perfekte ICCs konnten für die Hitzeschmerzschwelle (Hand) und die mechanische Schmerzempfindlichkeit (Rücken) errechnet werden. Substanzielle ICCs wurden für die Hitzeschmerzschwelle (Rücken), die Druckschmerzschwelle (Rücken), mechanische Schmerzempfindlichkeit (Hand) und Vibrationsschwelle (Rücken und Hand) beobachtet. Einige QST-Parameter, wie beispielsweise die Kälteerkennungsschwelle, wiesen niedrigere ICCs auf, aber auch eine sehr geringe Variabilität. Im Allgemeinen wiesen die QST-Messungen enge Grenzen der Übereinstimmung in den Bland-Altman-Diagrammen auf. Das standardisierte QST-Protokoll des DFNS ist für den Einsatz in Behandlungsstudien geeignet. Außerdem ist die Definition einer statistisch bedeutsamen Veränderung möglich, was eine Voraussetzung für den Einsatz von QST in klinischen Studien sowie in Langzeituntersuchungen ist.

3.3 Publikation III: Climate-related differences in temperature pain sensation? Comparison between a mid-European and a sub-Arctic sample.

Titel: Climate-related differences in temperature pain sensation? Comparison between a mid-European and a sub-Arctic sample.

Autoren: Helen Nothnagel, Thomas Weiss, Christian Puta, Gunnvald Kvarstein, Frauke Musial

Journal: *PeerJ* (IF 2.38). Eingereichtes Manuskript (eingereicht am 27. März 2021)

Eine mitteleuropäische (Deutschland) und eine norwegische (geographisch arktisch, klimatisch subarktisch) Forschergruppe kooperieren bei der Erforschung chronischer Schmerzen unter Verwendung des standardisierten QST-Protokolls des DFNS. Als Teil eines Qualitätssicherungsverfahrens wurden die QST-Messungen an der dominanten Hand (*dorsum manus*) zwischen den beiden Laboren verglichen. Die Daten beider Labore wurden im Rahmen von Studien erhoben, die Personen mit chronischen Nacken- oder Rückenschmerzen rekrutierten. Sowohl Teilnehmer, die als gesunde Kontrollen rekrutiert wurden, als auch Patienten wurden in die Qualitätssicherung eingeschlossen, wenn i) sie zum Zeitpunkt der Baseline-QST-Messung einen Wert > 40 für die Schmerzintensität auf einer numerischen Ratingskala von 0 bis 100 aufwiesen und ii) ihre QST-Messungen in Bezug auf die normativen Daten des DFNS innerhalb der Normdaten lagen. Die mitteleuropäische Stichprobe bestand aus 18 Probanden (42,8 ± 11,9 Jahre) und die subarktische Stichprobe aus 20 Probanden (41,1 ± 13,2 Jahre). Sowohl bei den mechanischen Tests als auch bei den thermischen Erkennungsschwellen gab es keinen signifikanten Gruppenunterschied. Allerdings wies die subarktische Stichprobe niedrigere Kälte- und Hitzeschmerzschwellen auf. Mögliche Störfaktoren, wie das Geschlecht, unterschiedliche Untersucher und ob die Studienteilnehmer als gesunde Kontrollen oder Patienten rekrutiert wurden, konnten ausgeschlossen werden. Die subarktische Stichprobe war empfindlicher gegenüber Kälte- und Hitzeschmerz, obwohl alle Teilnehmer innerhalb des Bereichs der QST-Normdaten lagen. Diese Ergebnisse legen nahe, dass mögliche regionale und klimabedingte Unterschiede einer Studienstichprobe bei der Bestimmung thermischer QST-Messwerte berücksichtigt werden müssen.

4 Abschlussdiskussion

Ziel der vorliegenden Dissertationsschrift ist die kritische Auseinandersetzung mit methodischen Herausforderungen der Messung chronischer Schmerzen bei der Durchführung klinischer Studien. Es werden drei Publikationen vorgestellt, die sich mit verschiedenen methodenkritischen Aspekten beschäftigen. Die Einordnung dieser Untersuchungen in ein konzeptuelles Framework ist in Abbildung 2 (Kapitel 1) dargestellt. Es werden drei wesentliche Aspekte (experimentelle und quasi-experimentelle Studiendesigns, Gütekriterien und ökologische Validität) behandelt und entsprechende Fragestellungen beantwortet. Zunächst wird die signifikante Abnahme der Schmerzintensität während der Rekrutierungsphase analysiert und diskutiert (Publikation I). Des Weiteren werden die Test-Retest-Reliabilität und Übereinstimmung der zur Messung der Schmerzempfindlichkeit angewandte QST-Methodik untersucht (Publikation II). In einer abschließenden Post-hoc Analyse werden die QST-Profile einer mitteleuropäischen und einer subarktischen Population verglichen (Publikation III).

4.1 Mögliche Fehlerquellen in der Rekrutierungsphase: Natural course of disease

Publikation I (Nothnagel et al., 2019) diskutiert die signifikante Abnahme der Schmerzintensität von 36,8 % während der Rekrutierungsphase einer klinischen Schmerzstudie, das heißt zwischen der ersten Kontaktaufnahme mit dem potenziellen Probanden am Telefon und der ersten Untersuchung. Es konnte eine signifikant negative Korrelation zwischen der Anzahl der Tage, die zwischen den beiden Messtagen lag, und der Veränderung der Schmerzintensität gefunden werden. Darüber hinaus ergab sich eine positive Korrelation zwischen der Veränderung der Schmerzintensität und der Schmerzintensität bei der ersten Kontaktaufnahme. Dies weist unter anderem darauf hin, dass die Abnahme der Schmerzintensität über die Zeit von der anfänglichen Schmerzintensität abhängig ist. Es konnten drei potenzielle Haupteinflüsse auf die klinisch relevante Abnahme der Schmerzintensität identifiziert werden: der natürliche Krankheitsverlauf, der Hawthorne Effekt und das Phänomen der Regression zum Mittelwert. Da alle drei Effekte nur schwer zu kontrollieren sind, können sie eine Verzerrung bei der

Erfassung von Messdaten in klinischen Studien zu Krankheiten mit einem schwankenden Symptommuster bedeuten. Alle aufgezeigten Phänomene stellen methodische Herausforderungen für klinische Studien zum chronischen Schmerz dar. Für eine Reihe von Schmerzforschern haben sie eine elementare Bedeutung und sind in weiteren klinischen Studien zum chronischen Schmerz zu berücksichtigen.

Auf der Basis der Ergebnisse der Publikation I können folgende Empfehlungen für weitere klinische Studien zum chronischen Schmerz gegeben werden:

Unter dem Aspekt des Hawthorne Effekts ist aufgrund der Tatsache, dass die Patienten an einer Studie teilnehmen dürfen, mit einer gewissen Abnahme der Schmerzintensität zu rechnen. Demnach sollten die Einschlusskriterien strenger eingeschränkt werden. Beispielsweise könnte das Einschlusskriterium der Schmerzintensität auf > 60 gesetzt werden.

Um den natürlichen Krankheitsverlauf bestmöglich kontrollieren zu können, ist ein enger Zeitrahmen, vor allem zwischen dem ersten Kontakt mit dem Patienten und der ersten Untersuchung, sinnvoll. Die Ergebnisse der Publikation I deuten darauf hin, dass die höchsten Werte in der Änderung der Schmerzintensität innerhalb der ersten 14 Tage eintraten. Demnach ist eine Zeitspanne zwischen dem ersten Kontakt mit dem potenziellen Patienten, dem Einschluss des Patienten und der ersten Untersuchung von wenigen Tagen bis maximal 1 Woche sinnvoll. Eine alternative Überlegung wäre, diese Zeitspanne weiter auszudehnen (länger als 14 Tage), um zumindest die anfänglichen Schwankungen der reportierten Schmerzintensität zu vermeiden.

Bei jeder klinischen Studie zum chronischen Schmerz, die die Schmerzintensität als wichtigen Outcome-Parameter verwendet, ist eine systematische Erfassung der Schmerzintensität unabdingbar. Bereits ab dem ersten Kontakt mit dem potenziellen Patienten einer Studie muss eine regelmäßige, beispielsweise tägliche, Dokumentation der Schmerzintensität erfolgen. Nur so lassen sich Symptomveränderungen im Laufe der Rekrutierungszeit erkennen. Darüber hinaus ist die Durchführung einer beobachtenden Pilotstudie vor Beginn einer Interventionsstudie ratsam, um den natürlichen Krankheitsverlauf in der Zielgruppe beobachten und analysieren zu können. Diese Daten ermöglichen dann einen systematischen Rekrutierungsansatz, weil Daten über ein von der Population abgeleitetes Einschlusskriterium, Informationen über den zeitlichen Verlauf des Symptoms sowie über die rekrutierbare Patientenpopulation zur Verfügung stünden.

4.2 Wie reproduzierbar ist die QST?

Der zweite Untersuchungsgegenstand (Publikation II) beinhaltet die detaillierte Analyse der Test-Retest-Reliabilität sowie der Übereinstimmung des standardisierten QST-Verfahrens über einen längeren, therapierelevanten Zeitraum (Nothnagel et al., 2017). Die Test-Retest-Reliabilitätsanalyse konnte zeigen, dass die meisten QST-Parameter über einen Zeitraum von 10 Wochen reliabel sind.

Eine nahezu perfekte Zuverlässigkeit konnte für die HPT sowohl am Rücken als auch an der dominanten Hand verifiziert werden. Vergleichbare Analysen anderer Studien bestätigen die Ergebnisse (Knutti et al., 2014; Marcuzzi et al., 2017). Wiederum andere Studien konnten nur eine mäßige bis moderate Zuverlässigkeit für HPT finden (Felix & Widerstrom-Noga, 2009; Moloney, 2011). Auch die in Publikation II untersuchte mäßige Zuverlässigkeit der CPT an Hand und Rücken trifft in der Literatur auf vergleichbare sowie gegensätzliche Ergebnisse (Felix, 2009; Moloney et al., 2011; Marcuzzi, 2017;). Wasner und Brock (2008) untersuchten die Reliabilität an drei verschiedenen Messzeitpunkten und zeigten eine deutlich bessere Zuverlässigkeit der CPT für ein Intervall von 1 Tag gegenüber dem Intervall von 21 Tagen. Diese Ergebnisse stimmen mit denen aus Publikation II überein. Für die thermischen Detektionsschwellen (CDT, WDT und TSL) wurden geringe bis moderate ICC-Werte über einen therapierelevanten Zeitraum analysiert. Moderate bis hohe ICC-Werte konnten sowohl am Rücken als auch an der dominanten Hand für alle mechanischen Parameter (mit Ausnahme der MDT am Rücken und WUR sowie MDT an der Hand) gefunden werden und lassen darauf schließen, dass die mechanischen QST-Parameter über die Zeit hinweg stabil sind. Frühere Studien berichteten über eine substanzielle bis nahezu perfekte Zuverlässigkeit für die PPT an Gesicht, Hals, Unterarm, Finger und Knie für ein Kurzzeitintervall (10 Minuten, 1 -3 Wochen) (Cathcart & Pritchard, 2006; Pigg et al., 2010; Wylde et al., 2011; Walton et al., 2014;). Andersen und Kollegen (2015) bestätigten eine hohe Zuverlässigkeit für MPT innerhalb einer Woche. Eine weitere Studie bestimmte eine nahezu perfekte Zuverlässigkeit für VDT und eine hohe Zuverlässigkeit für MDT für ein Kurzzeitintervall von 3 Wochen (Felix & Widerstrom-Noga, 2009).

Weitgehend zeigen die Daten der Publikation II moderate bis nahezu perfekte ICC-Werte sowohl für die thermischen als auch die mechanischen QST-Parameter am Rücken und an der dominanten Hand. Diese Ergebnisse finden sich überwiegend in der Literatur wieder. Abweichende Ergebnisse in der Literatur lassen sich möglicherweise auf die mangelnde

Standardisierung der QST-Methode, einschließlich der Verwendung verschiedener Geräte (Chong & Cros, 2004), der Rekrutierung unterschiedlicher Studienpopulationen, die selektive Testung beispielweise der thermischen QST-Parameter (Wasner & Brock, 2008; Moloney et al., 2011; Knutti et al., 2014) bzw. nur einiger ausgewählter QST-Parameter (Felix & Widerstrom-Noga, 2009; Wylde et al., 2011; Andersen, 2015) sowie verschiedener Zeitintervalle zwischen den Messzeitpunkten (Felix & Widerstrom-Noga, 2009; Moloney et al., 2011; Wylde et al., 2011; Knutti et al., 2014; Andersen et al., 2015) zurückführen.

In der Literatur werden die Begriffe Zuverlässigkeit (reliability) und Übereinstimmung (agreement) häufig synonym verwendet (Weir, 2005; De Vet et al., 2006; Kottner et al., 2011). Die daraus resultierenden unterschiedlichen statistischen Ansätze machen eine vergleichende Interpretation der Studienergebnisse nahezu unmöglich. Die Frage nach der Übereinstimmung zwischen wiederholten Messungen erfordert die Analyse, wie nah die Werte bei wiederholten Messungen beieinanderliegen (De Vet et al., 2006; Bartlett & Frost, 2008;). Bisher finden sich nur wenige Studien in der Literatur, die neben der Zuverlässigkeit auch die Übereinstimmung der QST-Parameter untersucht haben. Die Längsschnittuntersuchung zu Entwicklungsveränderungen der Somatosensorik von Hirschfeld et al. (2012) analysiert die Übereinstimmungsgrenzen für wiederholte QST-Messungen bei Kindern. Eine weitere Studie untersucht einige QST-Parameter 1 Jahr nach einer Brustkrebsoperation und berichtet über die Übereinstimmungsgrenzen für Unterarm, Bein und Brust (Andersen et al., 2015). Eigene Untersuchungsergebnisse zeigen sowohl kleinere als auch größere Übereinstimmungsgrenzen für verschiedene QST-Parameter. Verglichen mit den beiden oben genannten Studien entsprechen die eigenen analysierten Übereinstimmungsgrenzen einem kleineren bzw. einem gleich großen Bereich. Diese Ergebnisse scheinen überraschend, da zudem das in der Publikation II gewählte Zeitintervall zwischen den beiden Messzeitpunkten größer ist als in den Untersuchungen von Andersen et al. (2015) und Hirschfeld et al. (2012). Mögliche Gründe für den kleineren Bereich der Übereinstimmungsgrenzen der eigenen Untersuchung könnten in der untersuchten Studienpopulation (Patienten vs. gesunde Probanden) sowie in der standardisierten Durchführung der QST-Methode liegen.

Während die in Publikation II untersuchten ICC-Werte häufig nicht signifikant sind, weisen die identifizierten Übereinstimmungsgrenzen eher kleine Konfidenzintervalle auf. Dies scheint zunächst kontraintuitiv zu sein. Die Bland-Altman-Plots zeigen, dass viele der einzelnen QST-Werte innerhalb eines kleinen Bereiches liegen, was auf das Vorhandensein

eines Mittelwertes mit einer kleinen, aber zufälligen Variation schließen lässt. Der Mangel dieser systematischen Variation ist ausreichend, um das Fehlen einer signifikanten Korrelation aufzuklären. Dieser kleine Datenbereich einzelner QST-Parameter, insbesondere bei gesunden Probanden, macht diese Methode zu einem geeigneten Maß, um sowohl Abweichungen von der Norm bei Patientenpopulationen als auch durch Behandlungen induzierte Veränderungen erkennen zu können. Dies unterstreicht noch einmal die Bedeutung der QST innerhalb der Schmerzforschung sowie die gemeinsame Analyse der Zuverlässigkeit und Übereinstimmung.

Neben dem ICC ist auch der Standardfehler der Messung (SEM) als Maß für die Wiederholbarkeit der Messwerte von besonderer Bedeutung (De Vet et al., 2006; Kottner et al., 2011). Nach bisherigem Kenntnisstand findet sich in der Literatur nur eine Publikation, die die SEMs für verschiedene QST-Parameter (CPT, HPT, PPT und WUR) wiedergibt (Marcuzzi et al., 2017). Die in Publikation II untersuchten SEM-Werte sind mit den Ergebnissen von Marcuzzi et al. (2017) vergleichbar.

Nach derzeitigem Kenntnisstand sind die in Publikation II durchgeführten Untersuchungen hinsichtlich der Zuverlässigkeit und Übereinstimmung aller QST-Parameter nach standardisierten QST-Protokoll bei gesunden Probanden über einen therapierelevanten Zeitraum von 10 Wochen erstmalig in der Literatur dokumentiert. Die Kenntnis über die Variabilität der QST-Parameter ist eine notwendige Voraussetzung, um die durch Interventionen induzierten Veränderungen der Parameter beurteilen zu können. Die Ergebnisse der Studie von Nothnagel et al. (2017) legen nah, dass die QST sehr robust ist. Das heißt aber auch, dass die QST eben nicht sensitiv genug ist, um kleine Veränderungen über die Zeit erkennen zu können. Veränderungen über die Zeit müssen außerhalb der Übereinstimmungsgrenzen liegen, um überhaupt als solche erkannt werden zu können. Die Limits of agreement sind für die einzelnen QST-Parameter sehr unterschiedlich. Möglicherweise reichen einige QST-Parameter mit einem kleinen Übereinstimmungsbereich aus, um bereits kleine Veränderungen über die Zeit sehen zu können. In der klinischen Praxis könnte beispielsweise so das vollständige DFNS Protokoll zur initialen Diagnostik eingesetzt werden und für weitere Verlaufsmessungen würden einige, sensitive Parameter mit einem kleinen Bereich (z. B. HPT und PPT) ausreichen. Dies wiederum würde den zeitlichen Aufwand der gesamten QST-Testbatterie verringern und somit möglicherweise auch die Compliance der Probanden erhöhen.

Die Methode der QST wird seit vielen Jahren als Goldstandard in der klinischen Praxis zur Diagnostik neuropathischer Schmerzen eingesetzt. Prinzipiell ist die QST gut dafür geeignet, den Funktionszustand der primär-afferenten Fasersysteme zu erfassen. Dennoch sollte die QST sehr bewusst und reflektiert angewandt werden.

4.3 Einflüsse von Umwelt und Klima auf die Schmerzmessung

Um mögliche Unterschiede des QST-Profils einer mitteleuropäischen und einer subarktischen Population zu erkennen, wurde eine Post-hoc Analyse des QST-Protokolls vorgenommen (Publikation III: Nothnagel et al., 2021). Die Ergebnisse zeigen keine Unterschiede der mechanischen QST-Parameter (PPT, MPT, MPS, WUR, MDT und VDT) sowie der thermischen Detektionsschwellen (CDT, WDT und TSL) zwischen den analysierten Populationen. Dagegen weisen die thermischen Schmerzschwellen (CPT und HPT) Unterschiede zwischen den beiden Populationen auf. Im Vergleich zur mitteleuropäischen Stichprobe zeigt die subarktische Studienpopulation eine höhere Schmerzempfindlichkeit sowohl gegenüber Kälte als auch gegenüber Hitze.

Nach bisherigem Kenntnisstand finden sich in der Literatur keine vergleichbaren Publikationen, die diesen Ansatz untersucht haben. Möglicherweise lassen sich die analysierten Ergebnisse mit einem adaptierten und erlernten Verhalten der subarktischen Population gegenüber niedrigeren Temperaturen erklären. Diese Studienpopulation ist, gegenüber der mitteleuropäischen Population, über mehrere Monate im Jahr geringeren Durchschnittstemperaturen ausgesetzt, die möglicherweise als potenzielles Gesundheitsrisiko einzustufen sind.

Die in dieser Untersuchung resultierenden Ergebnisse zeigen einen sehr interessanten, neuartigen Befund. Die erhobenen Daten der beiden untersuchten Populationen befinden sich innerhalb der QST-Normdaten. Somit hat das Ergebnis der Studie für die Individualdiagnostik, für die die QST häufig verwendet wird, keine Auswirkungen. Demzufolge scheint das Ergebnis für die einzelne klinische Studie nicht relevant zu sein, jedoch umso mehr für die Vergleichbarkeit und die Verallgemeinerung der Studie. Besonders in Bezug auf multizentrische Studien zum chronischen Schmerz müssen Umweltbedingungen, wie beispielsweise klimatische Verhältnisse, berücksichtigt werden, um mögliche Verzerrungen der Ergebnisse zu vermeiden. Um den hier gezeigten Effekt

weiter systematisch zu untersuchen und klimabedingte Unterschiede der thermischen Schmerzschwellen differenziert aufzuklären, sind weitere Untersuchungen dringend erforderlich. Dies würde zu einer besseren Einordnung des Effekts in klinischen Studien zum chronischen Schmerz führen und mögliche Störfaktoren besser abschätzbar machen.

5 Fazit

Die in der vorliegenden Dissertation vorgestellten Untersuchungsergebnisse unterstreichen die methodenkritischen Aspekte und besonderen Herausforderungen klinischer Studien zum chronischen Schmerz bezogen auf die Schmerzmessung und -diagnostik. Publikation I analysiert die signifikante Abnahme der Schmerzintensität während der Rekrutierungsphase, noch vor Beginn der eigentlichen Intervention. Sowohl der natürliche Krankheitsverlauf, der Hawthorne Effekt als auch das Regression zum Mittelwert-Phänomen können die Ergebnisse einer klinischen Studie verzerren und sind damit besonders zu berücksichtigen. Studie II untersucht erstmals eine Test-Retest-Analyse des gesamten QST-Protokolls bei gesunden Probanden über einen therapierelevanten Zeitraum von 10 Wochen. Darüber hinaus werden Übereinstimmungsgrenzen (Limits of agreement) für alle QST-Parameter definiert. Diese Daten sind elementar, um die durch Interventionen induzierten Veränderungen der Parameter bewerten zu können und demzufolge beispielsweise den Erfolg einer Therapie bzw. einer Intervention bestimmen zu können. Publikation III zeigt einen Vergleich der thermischen QST-Parameter einer mitteleuropäischen und einer subarktischen Population. Diese vergleichende Analyse ist nach derzeitigem Kenntnisstand erstmalig. Die offene Fragestellung, inwieweit eine allgemeine Übertragbarkeit der DFNS QST-Referenzdaten für verschiedene Klimazonen möglich ist, muss jedoch durch weitere Forschungsarbeiten beantwortet werden. Möglicherweise sind noch weitere Einflüsse zu beachten, die bisher kaum bzw. gar nicht berücksichtigt wurden. Vor allem für die Vergleichbarkeit von Studien verschiedener Forschungseinrichtungen sowie für die Verallgemeinerung von Studienergebnissen ist das von entscheidender Bedeutung.

Die Kenntnis über mögliche Schwierigkeiten und methodenkritische Aspekte, die bei der Durchführung klinischer Studien zum chronischen Schmerz auftreten können, ist von großer Bedeutung. Die vorliegende Arbeit benennt die Herausforderungen und analysiert diese kritisch. Fragen „Wie kann man diese Herausforderungen lösen bzw. wie kann man damit umgehen?“ und „Welche Implikationen hat das für die klinische Praxis und Forschung?“ werden beantwortet. Zudem gibt diese Arbeit Empfehlungen sowohl für die klinische Praxis als auch die Forschung. Somit verfügt diese Arbeit über eine hohe Allgemeinrelevanz für klinische Studien zum chronischen Schmerz und leistet damit einen wichtigen Beitrag zur Schmerzforschung. Darüber hinaus dient die vorgelegte Dissertation als zentrale Vorarbeit

für zukünftige Studien, die die gezeigten Effekte und Herausforderungen weiter systematisch untersuchen werden.

Literaturverzeichnis

Aitken, L., Gallagher, R. & Madronio, C. (2003). Principles of recruitment and retention in clinical trials. *International journal of nursing practice*, *9* (6), 338-346.

Andersen, K.G., Kehlet, H. & Aasvang, E.K. (2015). Test-retest agreement and reliability of quantitative sensory testing 1 year after breast cancer surgery. *The Clinical journal of pain*, *31* (5), 393-403.

Aslaksen, P.M., Myrbakk, I.N., Høifødt, R.S. & Flaten, M.A. (2007). The effect of experimenter gender on autonomic and subjective responses to pain stimuli. *Pain, 129* (3), 260-268.

Backonja, M.M., Attal, N., Baron, R., Bouhassira, D., Drangholt, M., Dyck, P.J., Edwards, R.R., Freeman, R., Gracely, R., Haanpaa, M.H., Hansson, P., Hatem, S.M., Krumova, E.K., Jensen, T.S., Maier, C., Mick, G., Rice, A.S., Rolke, R., Treede, R.-D., Serra, J., Toelle, T., Tugnoli, V., Walk, D., Walalce, M.S., Ware, M., Yarnitsky, D. & Ziegler, D. (2013). Value of quantitative sensory testing in neurological and pain disorders: NeuPSIG consensus. *Pain, 154* (9), 1807-1819.

Baron, R. & Tölle, T.R. (2009). Schmerz und QST: "Messen, was messbar ist". *Der Schmerz, 23*, 5-6.

Bartlett, J.W. & Frost, C. (2008). Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, *31* (4), 466-475.

Bates, M.S., Edwards, W.T. & Anderson, K.O. (1993). Ethnocultural influences on variation in chronic pain perception. *Pain, 52*, 101-112.

Bellach, B.-M., Ellert, U. & Radoschewski, M. (2000). Epidemiologie des Schmerzes – Ergebnisse des Bundes-Gesundheitssurveys 1998. *Bundesgesundheitsblatt - Gesundheitsforschung – Gesundheitsschutz, 43* (6), 424-431.

Beubler, E. (2016). *Kompendium der medikamentösen Schmerztherapie. Wirkungen, Nebenwirkungen und Kombinationsmöglichkeiten* (6. Aufl.). Berlin: Springer.

Birbaumer, N.-P. & Schmidt, R.F. (2010). *Biologische Psychologie*. Heidelberg: Springer Medizin.

Birklein, F. & Sommer, C. (2013). Quantitative sensory testing - a tool for daily practice? *Nature Reviews Neurology*, *9* (9), 490-492.

Bland, J.M. & Altman, D.G. (2012). Agreed statistics: measurement method comparison. *Anesthesiology, 116* (1), 182-185.

Blankenburg, M., Boekens, H., Hechler, T., Maier, C., Krumova, E. , Scherens, A., Magerl, W., Aksu, F. & Zernikow, B. (2010). Reference value fur quantitative sensory testing in children and adolescents: Developmental and gender differences of somatosensory perception. *Pain, 149*, 76-88.

Bone, H.-G., Jaschik, M. & Pilkowski, K. (2010). Schmerz und Schmerztherapie. In L. Ullrich (Hrsg.), *Intensivpflege und Anästhesie* (S. 343-365). Stuttgart: Thieme.

Breivik, H., Collett, B., Ventafridda, V., Cohen, R. & Gallacher, D. (2006). Survey of chronic pain in Europe: Prevalence, impact on daily life, and treatment. *Eur J Pain, 10 (4), 287-287.*

Cathcart, S. & Pritchard, D. (2006). Reliability of pain threshold measurement in young adults. *The journal of headache and pain*, *7* (1), 21-26.

Chong, P.S.T. & Cros, D.P. (2004). Technology literature review: quantitative sensory testing. *Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine*, *2*(5), 734-747.

Dawson, A. & List, T. (2009). Comparison of pain thresholds and pain tolerance levels between Middle Easterners and Swedes and between genders. *Journal of oral rehabilitation, 36* (4), 271-278.

De Vet, H.C., Terwee, C.B., Knol, D.L. & Bouter, L.M. (2006). When to use agreement versus reliability measures. *Journal of clinical epidemiology*, *59* (10), 1033-1039.

Dieleman, J.P., Kerklaan, J., Huygen, F.J., Bouma, P.A. & Sturkenboom, M.C. (2008). Incidence rates and treatment of neuropathic pain conditions in the general population. *PAIN*, *137* (3), 681-688.

Downie, W.W., Leatham, P.A., Rhind, V.M., Wright, V., Branco, J.A. & Anderson, J.A. (1977). Studies with pain rating scales. *Annals of the Rheumatic Diseases, 1978* (37), 378-381.

Eberle, T., Doganci, B., Krämer, H.H., Geber, C., Fechir, M., Magerl, W. & Birklein, F. (2009). Warm and cold complex regional pain syndromes: differences beyond skin temperature? *Neurology*, *72* (6), 505-512.

Edwards, R.R. & Fillingim, R.B. (1999). Ethnic differences in thermal pain responses. *Psychosomatic medicine, 61* (3), 346-354.

Edwards, C.L., Fillingim, R.B. & Keefe, F. (2001). Race, ethnicity and pain. *Pain, 94* (2), 133-137.

Egloff, N., von Känel, R. & Egle, U.T. (2008). Weder Descartes noch Freud? Aktuelle Schmerzmodelle in der Psychosomatik. *Praxis*, *97* (10), 549-557.

Feine, J.S., Bushnell, M.C., Miron, D. & Duncan, G.H. (1991). Sex differences in the perception of noxious heat stimuli. *Pain, 44*, 255-262.

Felix, E.R. & Widerstrom-Noga, E.G. (2009). Reliability and validity of quantitative sensory testing in persons with spinal cord injury and neuropathic pain. *J Rehabil Res Dev*, *46* (1), 69-83.

Freynhagen, R., Rolke, R., Baron, R., Tölle, T.R., Rutjes, A.K., Schu, S. & Treede, R.D. (2008). Pseudoradicular and radicular low-back pain - a disease continuum rather than different entities? Answers from quantitative sensory testing. *PAIN*, *135* (1-2), 65-74.

Grond, S. & Radbruch, L. (2002). Systematik von Schmerzerkrankungen (akut, chronisch, Taxonomien). In H. Beck (Hrsg.), *Schmerztherapie* (S. 59-67). Stuttgart: Thieme.

Grouven, U., Bender, R., Ziegler, A. & Lange, S. (2007). Comparing methods of measurement. *Deutsche Medizinische Wochenschrift (1946)*, *132* (1), e69-73.

Gul, R.B. & Ali, P.A. (2010). Clinical trials: the challenge of recruitment and retention of participants. *Journal of clinical nursing*, *19* (1-2), 227-233.

Hansson, P., Backonja, M. & Bouhassira, D. (2007). Usefulness and limitations of quantitative sensory testing: clinical and research application in neuropathic pain states. *Pain*, *129* (3), 256-259.

Hirschfeld, G., Zernikow, B., Kraemer, N., Hechler, T., Aksu, F., Krumova, E., Maier, C., Magerl, W. & Blankenburg, M. (2012). Development of somatosensory perception in children: a longitudinal QST-study. *Neuropediatrics*, *43* (01), 10-16.

Huge, V., Lauchart, M., Förderreuther, S., Kaufhold, W., Valet, M., Azad, S.C., Beyer, A. & Magerl, W. (2008). Interaction of hyperalgesia and sensory loss in complex regional pain syndrome type I (CRPS I). *PloS one*, *3* (7), e2742.

Huppelsberg, J. & Walter, K. (2009). *Kurzlehrbuch Physiologie* (3. Aufl.). Stuttgart: Thieme.

Häuser, W., Schmutzer, G., Hinz, A., Hilbert, A. & Brähler, E. (2013). Prävalenz chronischer Schmerzen in Deutschland. *Der Schmerz*, *27* (1), 46-55.

Icenhour, A., Elsenbruch, S. & Benson, S. (2015). Biologische und psychosoziale Einflussfaktoren auf geschlechterbezogene Unterschiede beim Schmerz. *Gender: Zeitschrift für Geschlecht, Kultur und Gesellschaft*, *7* (2), 11-28.

Kállai, I., Barke, A. & Voss, U. (2004). The effects of experimenter characteristics on pain reports in women and men. *Pain, 112* (1-2), 142-147.

Kleist, P. (2006). Vier Effekte, Phänomene und Paradoxe in der Medizin. *Schweiz Medical Forum, 6* (46), 1023-1027.

Knutti, I.A., Suter, M.R. & Opsommer, E. (2014). Test–retest reliability of thermal quantitative sensory testing on two sites within the L5 dermatome of the lumbar spine and lower extremity. *Neuroscience letters*, *579*, 157-162.

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M. & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International journal of nursing studies*, *48* (6), 661-671.

Krumova, E.K., Geber, C., Westermann, A. & Maier, C. (2012). Neuropathic pain: is quantitative sensory testing helpful? *Current diabetes reports*, *12* (4), 393-402.

Kröner-Herwig, B. (2004). Schmerz - eine Gegenstandsbeschreibung. In H.-D. Basler (Hrsg.), *Psychologische Schmerztherapie* (S. 3-9). Berlin: Springer.

Lautenbacher, S. & Strian, F. (1991). Sex differences in pain and thermal sensitivity: The role of body size. *Perception & Psychophysics, 50* (2), 179-183.

Levine, F.M. & de Simone, L.L. (1991). The effects of experimenter gender of pain report in male and female subjects. *Pain, 44* (1), 69-72.

Maag, R., Binder, A., Maier, C., Scherens, A., Toelle, T., Treede, R.D. & Baron, R. (2008). Detection of a characteristic painful neuropathy in Fabry disease: a pilot study. *Pain Medicine*, *9* (8), 1217-1223.

Magerl, W., Krumova, E.K., Baron, R., Tölle, T., Treede, R.D., and Maier, C. (2010). Reference data for quantitative sensory testing (QST): refined stratification for age and a novel method for statistical comparison of group data. *Pain, 151* (3), 598-605.

Maier, C., Baron, R., Tölle, T.R., Binder, A., Birbaumer, N., Birklein, F., Gierthmühlen, J., Flor, H., Geber, C., Huge, V., Krumova, E.K., Landwehrmeyer, G.B., Magerl, W., Maihöfner, C., Richter, H., Rolke, R., Scherens, A., Schwarz, A., Sommer, C., Tronnier, V., Üçeyler, N., Valet, M., Wasner, G. & Treede, R.-D. (2010). Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS):

somatosensory abnormalities in 1236 patients with different neuropathic pain syndromes. *Pain, 150* (3), 439-450.

Marcuzzi, A., Wrigley, P.J., Dean, C.M., Adams, R. & Hush, J.M. (2017). The long-term reliability of static and dynamic quantitative sensory testing in healthy individuals. *Pain*, *158* (7), 1217-1223.

McCaffery, M., Beebe, A. & Latham, J. (1997). *Schmerz. Ein Handbuch für die Pflegepraxis.* Berlin: Ulstein Mosby.

Meh, D. & Denišlič, M. (1994). Quantitative assessment of thermal and pain sensitivity. *Journal of the Neurological Sciences, 127*, 164-169.

Melzack, R. & Wall, P.D. (1965). Pain mechanisms: a new theory. *Science*, *150* (3699), 971-979.

Moayedi, M. & Davis, K.D. (2013). Theories of pain: from specificity to gate control. *Journal of neurophysiology*, *109* (1), 5-12.

Moloney, N.A., Hall, T.M., O'Sullivan, T.C. & Doody, C.M. (2011). Reliability of thermal quantitative sensory testing of the hand in a cohort of young, healthy adults. *Muscle & nerve*, *44* (4), 547-552.

Mutschler, E., Schaible, H.-G. & Vaupel, P. (2007). *Anatomie, Physiologie, Pathophysiologie des Menschen* (6. Aufl.). Stuttgart: Wissenschaftliche Verlagsgesellschaft mbH.

Mücke, M., Cuhls, H., Radbruch, L., Baron, R., Maier, C., Tölle, T., Treede, R.-D. & Rolke, R. (2014). Quantitative sensorische Testung. *Der Schmerz*, *28* (6), 635-648.

Nilges, P. & Rief, W. (2016). Schmerzdiagnostik. *PiD-Psychotherapie im Dialog*, *17* (04), 34-38.

Nothnagel, H., Puta, C., Lehmann, T., Baumbach, P., Menard, M.B., Gabriel, B., Gabriel, H.H.W., Weiss, T. & Musial, F. (2017). How stable are quantitative sensory testing (QST) measurements over time? Report on 10-week reliability and agreement of results in healthy volunteers. *Journal of Pain Research, 10,* 2067-2078.

Nothnagel, H., Menard, M.B., Kvarstein, G., Norheim, A.J., Weiss, T., Puta, C., Mist, S.D. & Musial, F. (2019). Recruitment and inclusion procedures as “pain killers” in clinical trials? *Journal of pain research*, *12*, 2027-2037.

Nothnagel, H., Weiss, T., Puta, C., Kvarstein, G. & Musial, F. (2021). Climate-related differences in temperature pain sensation? Comparison between a mid-European and a sub-Arctic sample. *PeerJ* (zur Veröffentlichung eingereicht).

Pavlaković, G. & Petzke, F. (2010). The role of quantitative sensory testing in the evaluation of musculoskeletal pain conditions. *Current rheumatology reports*, *12* (6), 455-461.

Pfau, D.B., Krumova, E.K., Treede, R.D., Baron, R., Toelle, T., Birklein, F., Eich, W., Geber, C., Gerhardt, A., Weiss, T., Magerl, W. & Maier, C. (2014). Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): reference data for the trunk and application in patients with chronic postherpetic neuralgia. *Pain, 155* (5), 1002-1015.

Pigg, M., Baad-Hansen, L., Svensson, P., Drangsholt, M. & List, T. (2010). Reliability of intraoral quantitative sensory testing (QST). *Pain*, *148* (2), 220-226.

Probstfield, J.L. & Frye, R.L. (2011). Strategies for recruitment and retention of participants in clinical trials. *Jama*, *306* (16), 1798-1799.

Racine, M., Tousignant-Laflamme, Y., Kloda, L.A., Dion, D., Dupuis, G. & Choinière, M. (2012). A systematic literature review of 10 years of research on sex/gender and experimental pain perception–part 1: are there really differences between women and men? *Pain*, *153* (3), 602-618.

Radbruch, L. & Grond, S. (2002). Klinische Schmerzmessung. In H. Beck (Hrsg.), *Schmerztherapie* (S. 78-85). Stuttgart: Thieme.

Rolke, R., Baron, R., Maier, C., Tölle, T.R., Treede, R.-D., Beyer, A., Binder, A., Birbaumer, N., Birkleich, F., Bötefür, I.C., Braune, S., Flor, H., Huge, V., Klug, R., Landwehrmeyer, G.B., Magerl, W., Maihöfner, C., Rolko, C., Schaub, C., Scherens, A., Sprenger, T., Valet, M. & Wasserka, B. (2006a). Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): Standardized protocol and reference values. *Pain, 123*, 231-243.

Rolke, R., Magerl, W., Andrews Campell, K., Schalber, C., Caspari, S., Birklein, F. & Treede, R.-D. (2006b). Quantitative sensory testing: a comprehensive protocol for clinical trials. *European Journal of Pain, 10*, 77-88.

Röhrig, B., Du Prel, J.B. & Blettner, M. (2010). Studiendesign in der medizinischen Forschung—Teil 2 der Serie zur Bewertung wissenschaftlicher Publikationen. *Deutsche Zahnärztliche Zeitschrift*, *65* (9), 488.

Schaible, H.-G. (2010). Nozizeption und Schmerz. In R.F. Schmidt, F. Lang & M. Heckmann (Hrsg.), *Physiologie des Menschen: mit Pathophysiologie* (S. 299-313). Heidelberg: Springer Medizin.

Scherens, A., Maier, C., Haussleiter, I.S., Schwenkreis, P., Vlckova-Moravcova, E., Baron, R., & Sommer, C. (2009). Painful or painless lower limb dysesthesias are highly predictive of peripheral neuropathy: comparison of different diagnostic modalities. *European Journal of Pain*, *13* (7), 711-718.

Schulte am Esch, J. & Bause, H. (2011). *Anästhesie: Intensivmedizin, Notfallmedizin, Schmerztherapie* (4. Aufl.). Stuttgart: Thieme.

Spohn, D., Musial, F. & Rolke, R. (2013). Naturopathic reflex therapies for the treatment of chronic pain-part 2: quantitative sensory testing as a translational tool. *Complementary Medicine Research*, *20* (3), 225-230.

Treede, R.D., Kenshalo, D.R., Gracely, R.H. & Jones, A.K. (1999): The cortical representation of pain. *Pain, 79* (2-3), 105-111.

Treede, R.D., Rief, W., Barke, A., Aziz, Q., Bennett, M.I., Benoliel, R., Cohen, M., Evers, S., Finnerup, N.B., First, M.B., Giamberardino, M.A., Kaasa, S., Kosek, E., Lavand'homme, P., Nicholas, M., Perrot, S., Scholz, J., Schug, S., Smith, B.H., Svensson, P., Vlaeyen, J.W.S. & Wang, S.-J. (2015). A classification of chronic pain for ICD-11. *Pain*, *156* (6), 1003-1007.

Uddin, Z. & MacDermid, J.C. (2016). Quantitative sensory testing in chronic musculoskeletal pain. *Pain Medicine*, *17* (9), 1694-1703.

Walton, D.M., Levesque, L., Payne, M. & Schick, J. (2014). Clinical pressure pain threshold testing in neck pain: comparing protocols, responsiveness, and association with psychological variables. *Physical therapy*, *94* (6), 827-837.

Wasner, G.L. & Brock, J.A. (2008). Determinants of thermal pain thresholds in normal subjects. *Clinical Neurophysiology*, *119* (10), 2389-2395.

Weir, J.P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research*, *19* (1), 231-240.

Weisenberg, M. (1999). Cognitive aspects of pain. In P.D. Wall & R. Melzack (Ed.), *Textbook of pain* (pp. 345-357). Edinburgh: Churchill Livingstone.

Weiss, T. & Schaible, H.G. (2004). Strukturen der Nozizeption und der Schmerzverarbeitung. In F. van den Berg (Hrsg.). *Angewandte Physiologie. Schmerzen verstehen und beeinflussen* (S. 6-32). Stuttgart New York: Thieme.

Werner, M.U., Petersen, M.A. & Bischoff, J.M. (2013). Test-retest studies in quantitative sensory testing: a critical review. *Acta Anaesthesiol Scand,* 57 (8), 957-963.

Wylde, V., Palmer, S., Learmonth, I.D. & Dieppe, P. (2011). Test–retest reliability of Quantitative Sensory Testing in knee osteoarthritis and healthy participants. *Osteoarthritis and Cartilage*, *19* (6), 655-658.

Anhang

Publikation I

Journal of Pain Research

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Recruitment and inclusion procedures as "pain killers" in clinical trials?

This article was published in the following Dove Press journal:
Journal of Pain Research

H Nothnagel[1–3]
M Brown Menard[4,5]
G Kvarstein[6]
AJ Norheim[3]
T Weiss[7]
C Puta[1,8]
SD Mist[9]
F Musial[3]

[1]Department of Sports Medicine and Health Promotion, Friedrich Schiller University Jena, Jena, Germany; [2]Department of Psychiatry and Psychotherapy, Jena University Hospital, Jena, Germany; [3]Department of Community Medicine, The National Research Center for Complementary and Alternative Medicine (NAFKAM), UiT The Arctic University of Norway, Tromsø, Norway; [4]Crocker Institute, Kiawah Island, SC, USA; [5]School of Integrative Medicine and Health Sciences, Saybrook University, Oakland, CA, USA; [6]Pain clinic, University Hospital of Northern Norway, UiT The Arctic University of Norway, Tromsø, Norway; [7]Department of Biological and Clinical Psychology, Friedrich Schiller University, Jena, Germany; [8]Center for Interdisciplinary Prevention of Diseases related to Professional Activities, Friedrich Schiller University, Jena, Germany; [9]Oregon Health & Science University, Portland, OR, USA

Correspondence: H Nothnagel
Department of Sports Medicine and Health Promotion, Friedrich Schiller University Jena, Wöllnitzer Straße 42, Jena 07749, Germany
Tel +49 364 194 5647
Fax +49 3641 945652
Email helen.nothnagel@uni-jena.de

Background: Recruitment and inclusion procedures in clinical trials are time critical. This holds particularly true for studies investigating patients with fluctuating symptom patterns, like those with chronic neck pain. In a feasibility study on neck pain, we found a clinically relevant decrease in pain ratings within the recruitment period. This paper analyses the phenomenon and gives recommendations for recruitment procedures in clinical trials on pain.

Methods: Changes in pain intensity scores of 44 chronic neck pain patients (6 males and 36 females; mean age: 45.3±13.2 years) between the first telephone contact and baseline assessment were analyzed. Inclusion criterion was a mean pain intensity of ≥40 on a 0–100 numerical rating scale during the last three months. Statistical analyses were performed using ANOVA and parametric/non-parametric correlation coefficients.

Results: Average pain intensity score decreased significantly from 60.3±13.3 at telephone interview to 38.1±21.7 at baseline assessment. This represents a relative change of 36.8%. A weak but significant negative correlation was found between number of days between assessments and pain rating differences. There was a positive correlation between change of pain intensity and the pain level at the first contact, indicating that the decreased pain ratings over time were also dependent on the initial pain rating.

Conclusions: The clinically significant changes in pain intensity were weakly related to waiting time and moderately dependent on initial pain intensity, suggesting regression to the mean. The natural course of the disease and the Hawthorne effect are also discussed as contributing factors.

Keywords: chronic neck pain, Hawthorne effect, natural course of the disease, regression to the mean, clinical trial, recruitment

Introduction

Recruitment describes the selection process of participants/patients into a clinical trial, from the first communication to their enrolment, and is one of the most essential components in clinical research. A recruitment process can be complicated and time consuming, depending on the nature of the disease, the character of the inclusion/exclusion criteria, the type of intervention (single subject or group therapy), and the necessary procedures to check for those.

Complicated and time-consuming recruitment procedures are possibly acceptable for patients suffering from chronic diseases with a relatively stable time course of symptoms, but challenging for a clinical study recruiting patients suffering from pain syndromes with fluctuating symptom patterns such as musculoskeletal pain disorders. In diseases and syndromes with fluctuating pain, one might observe changes in pain ratings from inclusion to baseline which may obscure a potential treatment effect and contribute to high or early drop-out rates.

submit your manuscript | www.dovepress.com
DovePress
http://doi.org/10.2147/JPR.S204259

Generally, patients with fluctuating symptom patterns are usually contacted for participating in a potential study at times when the symptoms are particularly severe. If inclusion requires a time-consuming recruitment process, patients may experience reduced symptom severity at the time of enrolment, which will obscure the potential effect of the intervention investigated ("little or no pain – little or no gain"). In worst case, they may not fit the inclusion criteria anymore by the time the intervention begins.

Thus, the natural course of a disorder or a disease may play a significant role in clinical trials, even though it may be difficult to estimate its impact.[1–3] Only a few clinical trials provide a consistent report of the time period from the first communication with the study participant, baseline assessment, and details of the time course of the symptoms.

Performing a feasibility pilot study as a first step can help to identify and avoid threats to the implementation of a larger trial.[4,5] In a feasibility study on neck pain, our study group encountered a substantial decrease of pain intensity, between the telephone interview, as a first contact and the enrolment into the study. This change in pain ratings due to study routines and time schedule is relevant and can impose a threat to the main study itself.[6–12]

Therefore, the aim of this paper is to analyze the phenomenon of a clinically significant pain reduction after inclusion but before the beginning of the intervention and to investigate the contributing factors in order to avoid this phenomenon in larger trials on chronic non-specific neck pain.

Methods

General remarks

The data presented here are a part of a feasibility study, "Chronic, non-specific neck pain. Quantitative Sensory Testing (QST) as a tool for the investigation of massage and relaxation as interventions - a feasibility study" which aimed to explore study routines to prepare a clinical trial comparing clinical effects of massage and meditation as treatment interventions for chronic, non-specific neck pain including quantitative sensory testing as a biomarker. Although some details of the feasibility study are presented, our analyses will focus on the change in pain ratings during recruitment or the time period from the first contact to baseline assessment, just prior to the intervention.

Study participants

Study participants were recruited through local newspaper advertisements. During a telephone interview performed by one single person, they were screened for inclusion and exclusion criteria before referral to physical examination by a physician (A.J.N.).

Inclusion criteria were age between 21 and 75 years and non-specific neck pain persisting at least 3 months with a mean pain intensity of greater than or equal to 40 on a 100-point numerical rating scale (NRS, 0= "no pain at all", 100= "worst pain imaginable"). The patients were excluded if they suffered from a congenital deformation of the spine or had neurological symptoms, neuropathic pain, spine surgery less than 12 months prior to screening, or received transcutaneous electrical nerve stimulation (TENS), acupuncture, osteopathy, chiropractic maneuver, or infiltration in the area within 4 weeks prior to inclusion. With regard to medications, patients were excluded if they needed to take blood-thinning medicine, steroids, or strong pain medications, such as opioids. Patients were allowed to continue to use non-steroidal anti-inflammatory drugs (NSAID), if that was part of their current pain management. The use of medications was documented in a pain diary throughout the study.

The study was approved by the regional ethics committee (Regional committees for medical and health research ethics, REC North 2014/1105). Participants were informed about the purpose of the study, risk/benefit profile of the interventions and the study itself. All participants were free to withdraw from participation at any point if they wished to. All participants provided written informed consent prior to participation. The study was performed in accordance with the declaration of Helsinki.

Design

Figure 1 shows the study design of the feasibility study. After verification of eligibility of the study participants in telephone interview (T_0) and screening by study doctor, participants were included in the study. The study had a parallel design with two intervention arms; massage and relaxation. After baseline assessment (T_1), patients received five treatment sessions of either massage or relaxation, followed by post-intervention assessment (T_2), and follow-up assessment (T_3) 4 weeks later.

The analysis presented here focuses on the change of pain intensity indicated by study participants from the telephone interview (T_0) to the baseline assessment (T_1), before randomization into two study groups.

Primary outcome was the change of pain intensity. Pain intensity was recorded at all 4 assessments (T_0, T_1, T_2, and T_3) using the numerical rating scale (NRS, 0–100) at T_0 and the visual analog scale (VAS, 0–100) at T_1, T_2, and T_3. As

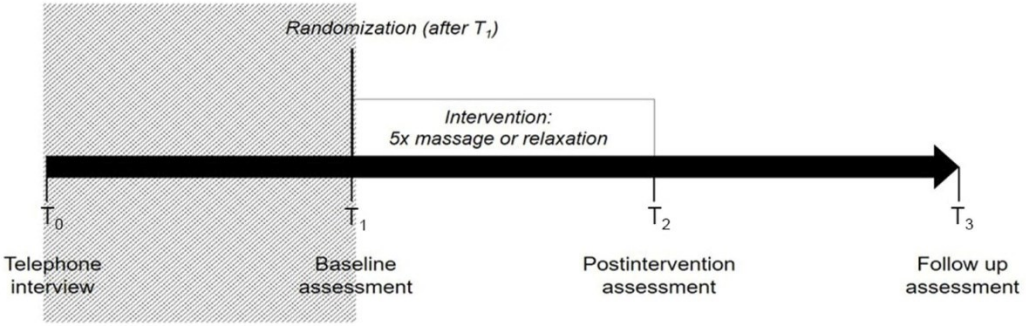


Figure 1 Study design of feasibility study across all assessments (T_0, T_1, T_2, and T_3). The analysis in the current study focuses on the time interval between the telephone interview (T_0) and the baseline assessment (T_1) presenting the gray structured square. The randomization into the groups (massage and relaxation) was performed after the baseline assessment (T_1).
Abbreviations: T_0, telephone interview; T_1, baseline assessment; T_2, post-intervention assessment; T_3, follow-up assessment.

secondary outcome, we performed a Quantitative Sensory Testing (QST) according to the standardized protocol establishing by the German Research Network on Neuropathic Pain (DFNS) at T_1, T_2, and T_3.[13]

Outcome measures

Pain

The analyses of the data presented in this paper focus on the differences of pain ratings given by the study participants between the telephone interview (T_0) and the baseline assessment (T_1). Pain intensity assessed during the telephone interview (T_0) was recorded on a numerical rating scale (NRS) graded from 0 to 100 (0= "no pain at all", 100= "worst pain imaginable"). Pain intensity assessed at baseline assessment prior to randomization (T_1) was recorded on a visual analog scale (VAS) graded from 0 mm to 100 mm (0 mm = "no pain at all", 100 mm = "worst pain imaginable"). Both, NRS and VAS scores were transferred in a common 0–100 pain rating scale to compare the pain ratings at both time points.

Expectation

Within the feasibility study, expectations regarding the anticipated treatment effects were measured at baseline assessment (T_1). The patients self-rated their expectations about the intervention on a VAS ranging from 0 mm to 100 mm (0 mm = "not effective at all", 100 mm = "highly effective") at the baseline assessment (T_1). Since expectations are seen as a part of a potential placebo effect within an interventional study, they are likewise relevant for the analyses presented in this paper.[14–16]

Statistical methods

Data for age, weight, height, BMI, pain intensity, and number of days between telephone interview (T_0) and baseline assessment (T_1) are presented as means with standard deviation (SD) and 95% confidence interval (95% CI of mean). We used Student *T*-test for paired samples to calculate the difference in pain intensity at T_0 and T_1. Normal distribution was tested using the Kolmogorov–Smirnov test.

For the statistical analyses of the change in pain intensity over time, the whole sample was further divided into two subgroups with regard to waiting time, ie, days between T_0 and T_1; group A: 1–14 days and group B: >14 days, based on a common waiting time after inclusion of about two weeks in clinical trials. The differences of pain ratings were analyzed with analysis of variance (ANOVA) using the differences of pain ratings between T_0 and T_1 as the within-subject factor and the grouping in two groups according to the days between the T_0 and T_1 as the between-subject factors. The pain rating (NRS 0–100) at T_0 was set as covariate in this model. No post hoc test was performed, since only two groups were compared.

Finally, three correlation analyses were performed to investigate a) the relationship between number of days between T_0 and T_1 and change in pain ratings, with the aim to identify a possible dependence on waiting time potentially related to the natural course of the disease, and b) the relationship between change in pain ratings between T_0 and T_1 and pain rating level at T_0, to identify regression to the mean, and c) the relationship between the difference of pain ratings between T_0 and T_1 and the expectation (0–100) about treatment effects at T_1, in order to investigate the impact of patient's expectations which may play a role as part of a placebo effect. Correlation analyses were conducted using Pearson's product–moment correlation coefficient for normally distributed parameters, and Kendall's Tau correlation coefficient

for non-normally distributed parameters. Correlation coefficients were interpreted as follows: |r| ≤0.29=negligible, 0.30–0.49=low, 0.50–0.69=moderate, 0.70–0.89 high, and >0.90=very high.

A *p*-value of <0.05 was considered statistically significant. Statistical analyses were performed using SPSS Statistics 22 (IBM Germany GmbH, Ehningen, Germany).

Results

Flow chart

A total of 85 potential patients were screened for eligibility by telephone interview (T_0), of these 51 were invited to the screening by the study physician (A.J.N.). A total of 44 patients fulfilled the inclusion criteria and agreed to participate in the study. The reasons for exclusion were pain intensity below 40 on a 0–100 pain rating scale at the time point of the clinical investigation (N=11), the localization of pain apart from the neck (N=6), and other ongoing, non-pharmacological treatments (N=2).

All patients underwent baseline assessment at (T_1). Two patients were considered as outlier according to waiting time between T_0 and T_1 and excluded from the statistical analyses. The outliers were defined as the values that were more than 1.5 x interquartile range beyond the 25th and 75th percentiles.[17]

Figure 2 shows a flow chart of patient flow throughout the feasibility study. The time points analyzed to determine possible changes in pain ratings during recruitment and inclusion are marked in black.

Sample characteristics

A total of 42 neck pain patients (45.3±13.2 years, mean±SD), 6 males and 36 females, were included in the statistical analyses. Average pain intensity and SD at T_0 was 60.3±13.3 and at T_1 38.1±21.7. The average time between T_0 and T_1 was 18.7±11.1 days. The expectation of the patients regarding the anticipated effectiveness of the treatment was 69.2±19.5 on a 0–100 VAS. Table 1 shows the demographic and clinical characteristics for the total sample.

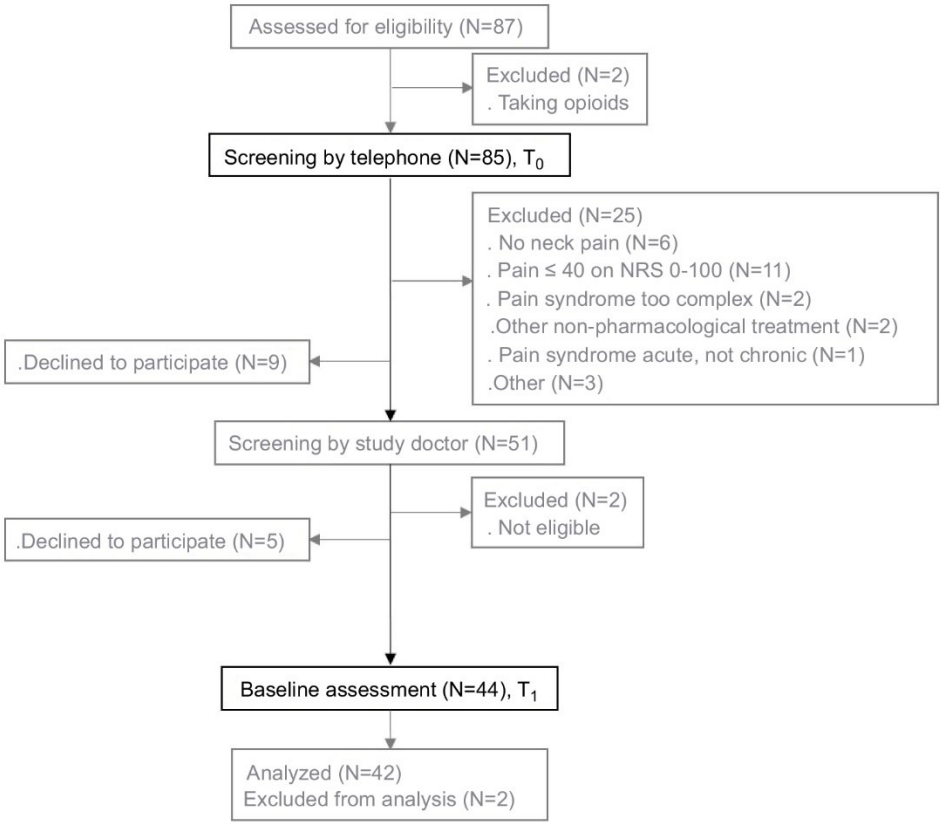


Figure 2 Flow chart of recruitment and progress of patients through study (the whole flow chart, gray). The analysis of the pain ratings during recruitment is based on pain intensity from the screening interview by telephone (T_0) and baseline assessment (T_1) (marked black in the flow chart).
Abbreviations: T_0, telephone interview; T_1, baseline assessment.

Table 1 Demographic and clinical characteristics of the total sample being analyzed

		Total
		N=42
Gender, N	**Male/female**	**6/36**
Age, years,[a]	Mean±SD (95% CI of mean)	45.3±13.2 (41.2–49.4)
Weight, kg,[a]	Mean±SD (95% CI of mean)	76.7±13.4 (72.6–80.9)
Height, cm,[a]	Mean±SD (95% CI of mean)	168.1±7.5 (165.7–170.4)
Body mass index, kg/m^{2},[a]	Mean±SD (95% CI of mean)	27.1±4.0 (25.9–28.3)
Pain intensity (NRS, 0–100) Telephone interview (T_0),[a]	Mean±SD (95% CI of mean)	60.3±13.3 (56.2–64.4)
Pain intensity (VAS, 0–100) Baseline assessment (T_1),[a]	Mean±SD (95% CI of mean)	38.1±21.7 (31.4–44.9)
Difference of pain intensity (0–100, T_0-T_1),[a]	Mean±SD (95% CI of mean) % *p*	22.2±25.4 (14.2–30.1) 36.8 0.000*
Pain medication (NSAID), N Baseline assessment (T_1)	Male/female	1/9
Days between Telephone interview (T_0) and Baseline assessment (T_1)	Mean±SD (95% CI of mean)	18.7±11.1 (15.3–22.2)
Expectation (0–100) Baseline assessment (T_1),[a]	Mean±SD (95% CI of mean)	69.2±19.5 (63.2–75.3)

Notes: Data are presented as mean±SD (95% CI of mean). [a]Normal distribution parameter by using the Kolmogorov–Smirnov test. NRS, 0–100 (0= "no pain", 100= "worst pain imaginable"); VAS, 0–100 (0= "no pain", 100= "worst pain imaginable").

Abbreviations: N, number of participants; SD, standard deviation; NRS, numerical rating scale; VAS, visual analog scale; NSAID, non-steroidal anti-inflammatory drugs.

Analyses of the pain ratings during recruitment and inclusion

Time course

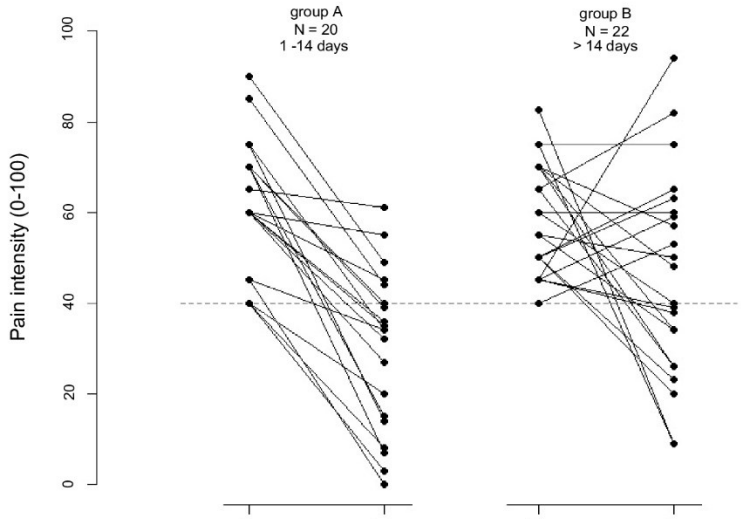
Of the whole sample (N=42), 34 participants showed a decrease in pain rating from T_0 to T_1. This was the case for all participants (N=20) who waited 1–14 days and for 14 out of 22 with a waiting time >14 days (N=22). Only 6 patients reported an increased pain intensity, while 2 patients reported similar values at T_0 and T_1 (see Figure 3). Figure 3 reveals that 24 participants exhibited a pain score of <40 at T_1 and were thus not fulfilling the main inclusion criterion to the feasibility study.

The average pain intensity at T_0 was 60.3±13.3 (0–100, mean±SD) and at T_1 38.1±21.7 (0–100, mean±SD). The difference in pain intensity between T_0 and T_1 corresponded 36.8% of the value at T_0.

The analysis of variance (ANOVA) for the pain ratings on T_0 and T_1 showed a significant main effect of the factor; days between T_0 and T_1 ($F(1,39)=6.17$; $p=0.017$; $\varepsilon=0.14$). Furthermore, we found a significant effect for the covariate pain rating (NRS, 0–100) at T_0 ($F(1,39)=14.14$; $p=0.001$; $\varepsilon=0.27$), indicating that the decrease in pain ratings over time was partly dependent on the initial pain rating at T_0.

Correlation of the change in pain scores and the initial pain scores at T_0

In order to further investigate the covariate pain rating at T_0, we conducted an additional correlation analysis including the change in pain ratings from T_0 to T_1 and the pain level at T_0. The analysis revealed a significant, although moderate positive correlation ($r=0.52$; $p<0.001$) supporting



Pain intensity (0-100)	T_0	T_1	T_0	T_1	T_0 (N=42)	T_1 (N=42)
mean ± SD	62.0 ± 14.5	29.0 ± 17.6	58.8 ± 12.2	45.6 ± 22.7	60.3 ± 13.3	38.1 ± 21.7
95 % CI of mean	55.2 - 68.8	21.7 - 38.1	53.3 - 64.2	35.6 - 55.7	56.2 - 64.4	31.4 - 44.9

Figure 3 Course of pain ratings (0–100) between the telephone interview (T_0) and baseline assessment (T_1). For all patients (N=42), both pain ratings were plotted before randomization. For the statistical analyses, the patients were classified into two groups depending on the waiting time between T_0 and T_1: group A patients with 1–14 days between T_0 and T_1 (N=20), group B patients with more than 14 days between T_0 and T_1 (N=22). The horizontal dashed line indicates the pain intensity criterion for inclusion into the feasibility study (greater than or equal to 40 on a 0–100 rating scale).
Abbreviations: T_0, telephone interview; T_1, baseline assessment.

the findings from the regression analysis above that the pain reduction from T_0 to T_1 was in part dependent on the initial pain rating at T_0 (see Figure 4).

Correlation of the change in pain scores with absolute waiting time

If the decrease in pain ratings is dependent on waiting time, then the difference in pain ratings between T_0 and T_1 should correlate with the absolute waiting time. The correlation analysis revealed a statistically significant negative but negligible correlation ($r = -0.25$; $p=0.020$) between number of days from T_0 to T_1 and the corresponding difference in pain ratings. This indicates that the decrease in pain ratings was not strongly dependent on waiting time (see Figure 5).

Correlation of the change in pain scores with expected treatment effects (expectation) at T_1

There was no significant correlation ($r=-0.01$; $p=0.948$) between the difference of pain ratings between T_0 and T_1, and the participant' expectation (0–100) about anticipated treatment effect (see Figure 6). Therefore, the decrease in pain ratings was not related to expectations about effectiveness of the interventions.

Discussion

The mean decrease in pain intensity from the first contact (T_0) until randomization (T_1) in our study was 36.8%. This is by definition above the cut off of 30% the minimal clinically important difference (MCID) and equivalent to a moderate pain reduction.[6,10,12,] Moreover, the effect was rather consistent, as 34 of the total sample (N=42) revealed a pain reduction during the waiting period.

Natural course of the disease

Neck pain fluctuates over time, and patients are probably most willing to participate in a study when their pain is temporarily more intense. Later, when the patient is to be randomized the pain levels may be lower due its natural course.[2,3]

The natural history of a disease describes its time course during the absence of an intervention.[1] This effect should have a similar influence on all study groups, if they are equally handled with regard to study procedures, and blinding and randomization are successful. However, the natural course of the disease may affect study outcomes substantially across groups, depending on the time course of the symptoms and the time interval between inclusion and intervention, and thus imposes a threat to the internal validity of a study.[1]

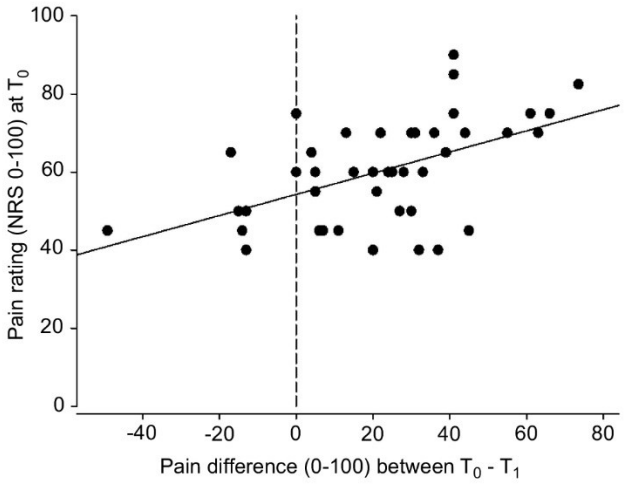


Figure 4 The scatter plot shows the relationship between pain changes from telephone interview (T_0) to baseline assessment (T_1) and the pain rating at T_0. For each patient (N=42), the change in pain rating (*x*-axis) is plotted against the pain level at T_0 (*y*-axis). The vertical dashed line represents a difference in pain rating of 0. Data points on the right illustrate patients with a decreasing pain rating to T_1, while data points on the left illustrate patients with an increasing pain rating to T_1.
Abbreviations: T_0, telephone interview; T_1, baseline assessment; NRS, numerical rating scale.

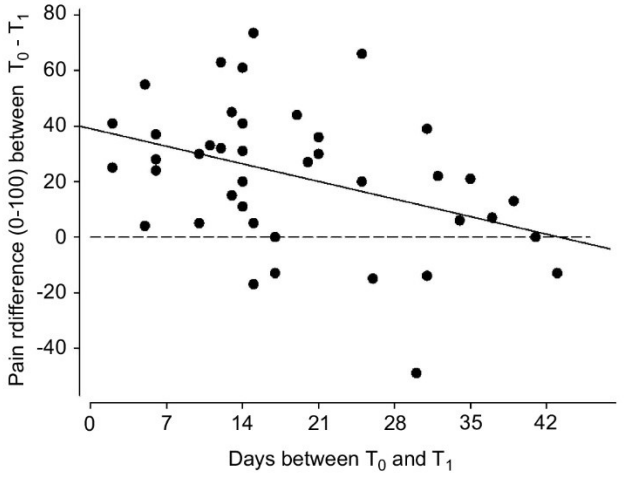


Figure 5 The scatter plot shows the relationship between waiting time until baseline assessment and differences in the pain intensity between the two measurement points. For each patient (N=42), the number of days between the telephone interview (T_0) and baseline assessment (T_1) (*x*-axis) are plotted against the differences in pain ratings from the telephone interview (T_0) and baseline assessment (T_1) (*y*-axis). The horizontal dashed line represents a change in pain rating (T_0 - T_1) of 0. Data points above the dashed line illustrate patients with a decreasing pain rating to T_1, while data points below the dashed line illustrate patients with an increasing pain rating to T_1.
Abbreviations: T_0, telephone interview; T_1, baseline assessment.

The significant correlation between waiting time and change in pain rating supports at first glance the assumption that a natural course of neck pain contributed to the effect. Nonetheless, the effect is not entirely consistent over time, as we also see less pain reduction with

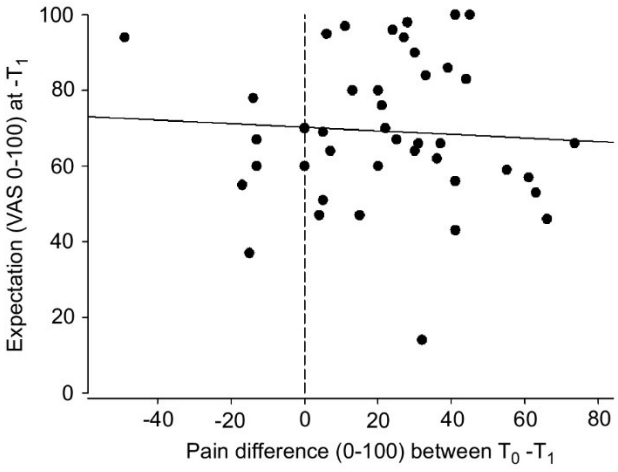


Figure 6 The scatter plot depicts the correlation between the differences in pain from the telephone interview (T_0) to baseline assessment (T_1), and the expectations of the study participants about the study outcome. For each patient (N=42), the differences (T_0-T_1) for the pain ratings (x-axis) are plotted against and the expectations at T_1 (*y*-axis). The vertical dashed line represents the differences of the pain ratings (T_0-T_1). The data points on the right illustrate patients with a decreasing pain rating up to T_1, data points on the left illustrate patients with an increasing pain rating up to T_1.
Abbreviations: T_0, telephone interview; T_1, baseline assessment; VAS, visual analog scale.

increased waiting time. A closer look at Figure 4 reveals negative values for six patients (left of the dotted vertical line). These were the patients who waited exceptionally long for T_1 (ranged from 15 to 43 days from T_0 to T_1). Thus, it is likely that these patients waited so long that they experienced an increase in pain ratings over time. Thus, the curve may in fact be more u-shaped as it appears in the correlational analysis, with an initial decrease in pain, followed by an increase over time.

Nonetheless, symptom fluctuation did not seem to be the only reason for the moderate pain reduction during the enrolment phase. The correlation between the change in pain rating and the initial pain rating at T_0 suggests that the change in pain rating over time is also dependent on initial pain intensity at T_0. A fact that could support the hypothesis that participants are more likely to enter a study when their pain is particularly intense.

Regression to the mean

Another highly relevant phenomenon with regard to fluctuating symptom patterns is the so-called "regression to the mean".[18–20] Regression to the mean is a purely statistical phenomenon, describing the general tendency for extreme values to converge towards a middle level.[21] The more extreme the initial value is, the higher the potential for such regression will be.[22]

For fluctuating pain syndromes, this implies that a high pain intensity prior to examination will to some extent, in itself, predict the following decrease. Taken together, the fact that it is necessary to define a reasonably high initial value as inclusion criterion in pain trials implies that regression to the mean is likely to occur. Regression to the mean can affect any investigation where the response to treatment is classified relative to initial values.[23]

In conclusion, both the natural course of the pain syndrome and regression to the mean may have influenced the decrease in pain intensity in our study.[1,24] A predefined inclusion criterion with high pain intensity will lead to a data structure, where regression to the mean is likely to occur.

Possible placebo effects and expectation

Expectation-dependent placebo effects are an inherent part of all clinical interventions and must be taken into account. It is therefore a standard baseline procedure in clinical trials to include questions about expectations about the treatment effects. This was therefore done at T_1 in the planned feasibility study. Although we did not assess expectation at T_0, we assumed that by the fact that the study participants showed up at T_1 to enter the interventional part of the study that their expectations about the effectiveness of the treatment were at least as high at T_1. Therefore, even though the correlation between the participant's expectation at T_1 and the difference in pain ratings between T_0 and T_1 represents a "backward analysis", it still provides information on how expectation influenced the difference in pain ratings.[14–16]

The correlation analysis revealed that there was no influence of treatment effect related expectations on the difference in pain ratings. Figure 6 shows the reason for that: treatment effect expectations were uniformly high (69.2±19.5, Table 1) at T_1 so that there was not enough variation in expectations to correlate with the change in pain ratings. In conclusion, even though we should consider expectation-dependent placebo effects as part of the treatment effect in the feasibility study (not presented here), we find it unlikely, that the expectation-dependent placebo effects played a role for the reduction in pain ratings within the waiting period.

The Hawthorne effect

Another, critical issue with regard to fluctuations in outcome measures may be the so-called Hawthorne effect. This effect describes the phenomenon that a person may change her or his behavior, experiences, emotions, etc., when becoming a study participant.[25–28] The phenomenon is interpreted as a type of reactivity to the situation, where a person is being systematically investigated and "observed". A possible reason for this effect may be increased attention to factors that are related to the study outcomes.

Thus, independent of the natural course of the disease, the inclusion procedure and the enrolment into a study may, in itself, have a major impact on the main study outcomes. Even though this influence can be expected to be equally distributed among the study groups,[29–32] the Hawthorne effect is difficult to control in experimental as well as, placebo and non-treatment control groups.

Limitations

The data presented are part of the feasibility study on how to integrate the standardized protocol of DFNS QST protocol into a clinical trial on the effects of massage and meditation for chronic, non-specific neck pain. However, the study was not planned to directly assess recruitment and enrolment-related problems. Thus, a more differentiated design, directly aiming at possible threats to the internal validity of outcomes in pain trials during the recruitment and enrolment phase, would have been more appropriate. Nonetheless, we assume that the challenges to a clinical study on pain presented here are of relevance to other pain researchers and have to be regarded in further clinical trials.

The use of two different pain rating scales in this study is another limitation. During the screening (telephone interview, T_0), patients assessed pain intensity with a numeric rating scale (NRS, 0–100) while a visual analog scale (VAS, 0–100) was used at baseline assessment (T_1). The reason for this incongruence is that the NRS is easier to handle in a telephone interview situation.

Although it has been shown that NRS and VAS ratings correlate significantly, VAS scores have a tendency to reveal lower ratings compared to NRS scores. Thus, it is not unproblematic to use these two scales interchangeably when assessing self-reported pain.[33,34] In our study, the fact that NRS scores seem to reveal higher pain ratings than VAS scores means that the pain scores measured at T_0 possibly overestimate pain compared to the VAS pain measures at T_1. Ultimately, this overestimation would contribute to higher pain values at the first measurement point (T_0) and could thus lead to an overestimation of the decrease in pain ratings from T_0 to T_1. In conclusion, it would be important to estimate whether the decrease in

pain ratings from T_0 to T_1 is within or outside the range of the difference in pain ratings between the NRS and VAS ratings.

According to the literature, the differences between NRS and VAS means were |0.52| cm for the initial assessment and |0.86| cm across all assessments in a study conducted among teenagers (mean age 14.7±3.1 years). Moreover, in this study, the differences between NRS and VAS ratings decreased with increasing pain severity across all assessments.[34] In another study conducted with adults, the differences between NRS and VAS ranged between |2.30| cm and |1.30| cm.[34]

In our study, the difference between the telephone interview (T_0) and the baseline assessment (T_1) was 60.3±13.3 and 38.1±21.7 representing a change of 36.8% and a distance of |22.2| mm or |2.22| cm between the NRS at T_0 and the VAS at T_1. According to the data on children provided by Myrvik (2015) this difference of 2.22 cm would be outside the range of differences *between* the two scales, but within the range according to the data provided Holdgate (2003), even though at the outer limit of the interval.

In conclusion, the fact that the pain ratings measured with the NRS tend to be higher than those measured with the VAS and the fact that the NRS was used at T_0 might have contributed to systematically higher pain ratings at T_0 in this study. It may have thus contributed to larger decreases in pain ratings from T_0 to T_1.[34] However, since we expected regression to the mean to play a role in this analysis, our statistical approach already controls for the dependency of the differences between T_0 and T_1 on the initial values at T_0. In our ANOVA model, the initial values were used as covariate, and thus, the influence of the initial values on the total effect is extracted. Indeed, there is a significant effect of the covariate (pain rating NRS at T_0), nonetheless, the factor days between T_0 and T_1 remained significant and thus non-neglectible. It is impossible to estimate how much of the difference between NRS and VAS contributed to the effect of the increased initial values (pain ratings at T_0); however, the ANCOVA reveals a significant effect of waiting time beyond a significant effect of the covariate. Thus, even though the initial values at T_0 played an important role for the total effect, there still remains substantial support for a decrease in pain ratings during the waiting time between recruitment and inclusion.

Further directions and recommendations

The Hawthorne effect predicts that a certain reduction in pain has to be expected simply due to the fact that patients participate in a study. Consequently, inclusion criteria should possibly be strict, eg, pain intensity >60. There are several challenges related to this: i) The total available study population will be significantly smaller and consequently, recruitment may be more difficult; ii) Such a high pain intensity may not reflect the majority of the patients and will thus restrict the generalization of the study results (external validity); iii) A high pain rating as inclusion criterion will also make the main outcome more vulnerable to distribution-related phenomena, such as regression to the mean.

While the Hawthorne effect is uncontrollable, the impact of the natural course of the disease could be controlled by keeping the time frame as tight as possible. In our study on neck pain, the main reduction in pain intensity occurred within 2 weeks and thus a time frame of a few days would be more advisable.

However, another methodological approach to the natural course of the disease and regression to the mean could be to wait more than 14 days in order to avoid the initial fluctuations in pain ratings (see Figures 3 and 4 where six study participants already show an increase in pain ratings over time). The benefits of such a design would be that it is easier to pick up a defined treatment effect while the disadvantages are that the study population may not represent typical pain patients.

Moreover, the main outcome measure of a study, in this case the pain intensity, should be recorded systematically after the first contact (eg, telephone interview) throughout the recruitment and enrolment. This allows for better documentation of symptom changes over time. Another possibility would be to conduct an observational pilot study before an interventional trial to observe the natural course of the disease in the target group. Such data would allow for a more systematic recruitment approach since a population-derived inclusion criterion, information about the time course of the symptom, and the recruitable patient population would be available.

Conclusion

Three main potential influences on a clinically relevant decrease in pain scores during the enrolment phase of a study were identified; the natural course of the disease, regression to the mean, and the Hawthorne effect. They can all impose a threat to a clinical trial of diseases with a fluctuating symptom pattern because they are difficult to control. Clinical trial methodology, by defining a primary inclusion criterion related to the primary outcome measure, comprises a risk of a regression to the mean in spontaneously fluctuating diseases. Awareness of this

phenomenon can contribute to better routines for clinical studies accounting for a change in the main outcome measure during enrolment and recruitment.

Acknowledgments

We thank Brit Johanne Dragest for her indispensable assistance in the organization and coordination of the study. Tone Westergård and Elke Klemowicz are acknowledged for their practical work and engagement in all massage and relaxation interventions. We are indebted to all voluntary patients who participated in the study for their consent and cooperation.

Author Contributions

All authors contributed to data analysis, drafting and revising the article, gave final approval of the vision to be published, and agree to be accountable for all aspects of the work.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Jekel JF, Katz DL, Elmore JG, Wild DMG. *Epidemiology, Biostatistics, and Preventive Medicine*. Philadelphia: Saunders; 2007.
2. Stub T, Salamonsen A, Alræk T. Safety of treatment provided by homeopaths. *Eur J Integr Med*. 2016;8:30–31. doi:10.1016/j.eujim.2016.08.074
3. Stub T, Salamonsen A, Kristoffersen A, Musial F. How to handle worsening of condition during treatment – risk assessment in homeopathic practice. *Forsch Komplementmed*. 2015;22(1):30–35. doi:10.1159/000377644
4. Arain M, Campell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Med Res Methodol*. 2010;10(1):67. doi:10.1186/1471-2288-10-67
5. Thabane L, Ma J, Chu R, et al. A tutorial on pilot studies: the what, why and how. *BMC Med Res Methodol*. 2010;10(1):1. doi:10.1186/1471-2288-10-1
6. Dworkin RH, Turk DC, Wyrwich KW, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J Pain*. 2008;9(2):105–121. doi:10.1016/j.jpain.2007.09.005
7. Farrar JT, Young JP Jr, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain*. 2001;94(2):149–158.
8. Farrar JT, Pritchett YL, Robinson M, Prakash A, Chappell A. The clinical importance of changes in the 0 to 10 numeric rating scale trials of duloxetine in pain disorders. *J Pain*. 2010;11(2):109–118. doi:10.1016/j.jpain.2009.06.007
9. Jensen MP, Chen C, Brugger AM. Interpretation of visual analog scale ratings and change scores: a reanalysis of two clinical trials of post-operative pain. *J Pain*. 2003;4(7):407–414.
10. Mannion AF, Balagué F, Pellise F, Cedraschi C. Pain measurement in patients with low back pain. *Nat Clin Pract Rheumatol*. 2007;3(11):610–618. doi:10.1038/ncprheum0646
11. Morley S, Williams A, Hussain S. Estimating the clinical effectiveness of cognitive behavioural therapy in the clinic: evaluation of a CBT informed pain management programme. *Pain*. 2008;137(3):670–680. doi:10.1016/j.pain.2008.02.025
12. Turner JA, Turk DC. The significance of clinical significance. *Pain*. 2008;137(3):467–468. doi:10.1016/j.pain.2008.03.029
13. 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. doi:10.1016/j.ejpain.2005.02.003
14. Enck P, Benedetti F, Schedlowski M. New insights into the placebo and nocebo responses. *Neuron*. 2008;59(2):195–206. doi:10.1016/j.neuron.2008.06.030
15. Horing B, Weimer K, Muth ER, Enck P. Prediction of placebo responses: a systematic review of the literature. *Front Psychol*. 2014;5:1079. doi:10.3389/fpsyg.2014.01079
16. Musial F, Klosterhalfen S, Enck P. Placebo responses in patients with gastrointestinal disorders. *World J Gastroenterol*. 2007;13(25):3425. doi:10.3748/wjg.v13.i25.3425
17. Bortz J. *Statistik Für Human- Und Sozialwissenschaftler*. Heidelberg: Springer; 2010.
18. Bland JM, Altman DG. Statistics notes: some examples of regression towards the mean. *Bmj*. 1994;309:780. doi:10.1136/bmj.309.6957.780
19. Davis CE. The effect of regression to the mean in epidemiologic and clinical studies. *Am J Epidemiol*. 1976;104(5):493–498. doi:10.1093/oxfordjournals.aje.a112321
20. Yudkin PL. How to deal with regression to the mean in intervention studies. *Lancet*. 1996;347:241–243.
21. Maher M, Mountain L. The sensitivity of estimates of regression to the mean. *Accid Anal Prev*. 2009;41(4):861–868. doi:10.1016/j.aap.2009.04.020
22. Morton V, Torgerson DJ. Effect of regression to the mean on decision making in health care. *Bmj*. 2003;326(7398):1083–1084. doi:10.1136/bmj.326.7398.1083
23. Shephard RJ. Regression to the mean. A threat to exercise science? *Sports Med*. 2003;33(8):575–584. doi:10.2165/00007256-200333080-00003
24. Barnett AG, Van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol*. 2005;34(1):215–220. doi:10.1093/ije/dyh299
25. McCambridge J, Witton J, Elbourne DR. Systematic review of the hawthorne effect: new concepts are needed to study research participation effects. *J Clin Epidemiol*. 2014;67(3):267–277. doi:10.1016/j.jclinepi.2013.08.015
26. McCarney R, Warner J, Iliffe S, Van Haselen R, Griffin M, Fisher P. The hawthorne effect: a randomized, controlled trial. *BMC Med Res Methodol*. 2007;7(1):30.
27. Parsons HM. What happened at hawthorne?: new evidence suggests the hawthorne effect resulted from operant reinforcement contingencies. *Science*. 1974;183(4128):922–932. doi:10.1126/science.183.4128.922
28. Wickström G, Bendix T. The "hawthorne effect" – what did the original hawthorne studies actually show? *Scand J Work Environ Health*. 2000;26(4):363–367.
29. Adair JG. The hawthorne effect: a reconsideration of the methodological artifact. *J Appl Psychol*. 1984;69(2):334–345. doi:10.1037/0021-9010.69.2.334
30. De Amici D, Klersy C, Ramajoli F, Brustia L, Politi P. Impact of the hawthorne effect in a longitudinal clinical study: the case of anesthesia. *Control Clin Trials*. 2000;21(2):103–114.

31. Braunholtz DA, Edwards SJ, Lilford RJ. Are randomized clinical trials good for us (in the short term)? Evidence for a "trial effect". *J Clin Epidemiol*. 2001;54(3):217–224. doi:10.1016/S0895-4356(00)00305-X
32. Leonard KL. Is patient satisfaction sensitive to changes in the quality of care? An exploitation of the hawthorne effect. *J Health Econ*. 2008;27(2):444–459. doi:10.1016/j.jhealeco.2007.07.004
33. Holdgate A, Asha S, Craig J, Thompson J. Comparison of a verbal numeric rating scale with the visual analogue scale for the measurement of acute pain. *Emerg Med*. 2003;15(5-6):441–446. doi:10.1046/j.1442-2026.2003.00499.x
34. Myrvik MP, Drendel AL, Brandow AM, Yan K, Hoffmann RG, Panepinto JA. A comparison of pain assessment measures in pediatric sickle cell disease: visual analog scale versus numeric rating scale. *J Pediatr Hematol Oncol*. 2015;37(3):190–194. doi:10.1097/MPH.0000000000000306

Publikation II

Journal of Pain Research

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

How stable are quantitative sensory testing measurements over time? Report on 10-week reliability and agreement of results in healthy volunteers

Helen Nothnagel[1,2,*]
Christian Puta[1,3,*]
Thomas Lehmann[4]
Philipp Baumbach[5]
Martha B Menard[6,7]
Brunhild Gabriel[1]
Holger H W Gabriel[1]
Thomas Weiss[8]
Frauke Musial[2]

[1]Department of Sports Medicine and Health Promotion, Friedrich Schiller University, Jena, Germany; [2]Department of Community Medicine, National Research Center in Complementary and Alternative Medicine, UiT, The Arctic University of Norway, Tromsø, Norway; [3]Center for Interdisciplinary Prevention of Diseases Related to Professional Activities, [4]Department of Medical Statistics, Computer Sciences and Documentation, Friedrich Schiller University, [5]Department of Anesthesiology and Intensive Care Medicine, University Hospital Jena, Germany; [6]Crocker Institute, Kiawah Island, SC, [7]School of Integrative Medicine and Health Sciences, Saybrook University, Oakland, CA, USA; [8]Department of Biological and Clinical Psychology, Friedrich Schiller University, Jena, Germany

*These authors contributed equally to this work

Correspondence: Helen Nothnagel
Department of Sports Medicine and Health Promotion, Friedrich Schiller University, Wöllnitzer Straße 42, 07749 Jena, Germany
Tel +49 3641 94 5650
Fax +49 3641 94 5652
Email helen.nothnagel@uni-jena.de

Background: Quantitative sensory testing (QST) is a diagnostic tool for the assessment of the somatosensory system. To establish QST as an outcome measure for clinical trials, the question of how similar the measurements are over time is crucial. Therefore, long-term reliability and limits of agreement of the standardized QST protocol of the German Research Network on Neuropathic Pain were tested.

Methods: QST on the lower back and hand dorsum (dominant hand) were assessed twice in 22 healthy volunteers (10 males and 12 females; mean age: 46.6±13.0 years), with sessions separated by 10.0±2.9 weeks. All measurements were performed by one investigator. To investigate long-term reliability and agreement of QST, differences between the two measurements, correlation coefficients, intraclass correlation coefficients (ICCs), Bland–Altman plots (limits of agreement), and standard error of measurement were used.

Results: Most parameters of the QST were reliable over 10 weeks in healthy volunteers: Almost-perfect ICCs were observed for heat pain threshold (hand) and mechanical pain sensitivity (back). Substantial ICCs were observed for heat pain threshold (back), pressure pain threshold (back), mechanical pain sensitivity (hand), and vibration detection threshold (back and hand). Some QST parameters, such as cold detection threshold, exhibited low ICCs, but also very low variability. Generally, QST measures exhibited narrow limits of agreement in the Bland–Altman plots.

Conclusion: The standardized QST protocol of the German Research Network on Neuropathic Pain is feasible to be used in treatment trials. Moreover, defining a statistically meaningful change is possible, which is a prerequisite for the use of QST in clinical trials as well as in long-term investigations of disease progression.

Keywords: QST, healthy volunteers, test–retest reliability, intraclass correlations, Bland–Altman plot, limits of agreement, standard error of measurement, minimum detectable difference

Introduction

Quantitative sensory testing (QST) investigates the submodalities of the somatosensory system, such as temperature, touch, vibration, and pain. It provides information on the state of peripheral sensory nerves, as well as pain perception and central sensitization. The method allows for the evaluation of the functional status of the small (Aδ, C) and large (Aβ) fiber sensor systems.[1,2]

Over the past decade, QST has achieved a unique position within the field of pain diagnostics.[3–9] However, heterogeneity of the protocols remains a challenge. Recently, a highly standardized QST protocol was established by the German Research Network

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/JPR.S137391

on Neuropathic Pain (DFNS), including a reference database with age- and gender-matched normative data from healthy volunteers for face, hand, and foot.[2,10,11] Additional reference values for the back were established.[12] Consequently, QST has been proven to discriminate between pathological states reliably from among those found in a normal population.[13,14] Because of its unique discriminative properties, it can be assumed that QST is likewise able to detect changes over time. This is obviously relevant regarding the detection of pathological changes and functional deterioration, such as those observed in diabetes, and could also be relevant for quantification of treatment effects in a clinical trial on pain. However, a crucial precondition for the usefulness of QST as an outcome measure in clinical trials is that repeated measures of QST are principally stable over time.

While the reliability of QST has been investigated before, available data only include several QST parameters and submodalities measured within a few days up to 1 month.[15–25] These studies found generally high test–retest reliabilities. The test–retest reliability of the complete QST protocol of the DFNS has only been investigated in a few studies.[26–29] Geber et al[29] analyzed 60 QST profiles from patients with sensory disturbances in a multicenter study. They found a high test–retest reliability for all parameters over a period of 2 days. Moreover, Pigg et al[28] found acceptable intraoral reliabilities between 6 and 21 days. Hirschfeld et al[27] conducted QST in a cohort of children and adolescents over a period of 15 months and found a systematic decrease in pain sensitivity, most likely related to maturation. However, information about its stability over time under most ideal circumstances is required in order to establish QST as a valid instrument for clinical trials. In our understanding, optimal circumstances would include an investigation over several weeks with a highly standardized protocol in healthy, pain-free subjects. To our knowledge, there is only one paper available addressing this point.[30]

A particular, methodological problem of repeated measures is the fact that data can correlate highly yet may not present a stable repetition of a measurement. As Bland and Altman[31] point out: "If one measurement is always twice as big as the other, they are highly correlated, but they do not agree." They suggest statistically defining the limits of agreement (LoAs) between two measurements. Moreover, it is most likely the agreement between two measurements that determines the usefulness for detecting meaningful changes over time for quantifying a treatment effect or documenting pathological deterioration. To date, no information about the LoAs under optimized conditions is available for QST. Therefore, the present study aims to determine the long-term reliability and agreement for a period of 10 weeks in adult healthy volunteers using the standardized QST protocol of the DFNS.

Methods

Participants

Twenty-two healthy volunteers, 10 males (38.2±13.1 years, mean ± SD) and 12 females (54.2±6.8 years, mean ± SD), were investigated between March 2013 and September 2014 (Table 1). All participants were free of pain (numerical rating scale, NRS =0). In addition, they self-reported no use of any pain medication for at least 48 hours before the first QST session. The participants were volunteers recruited from a participant database and student population of the Department of Sports Medicine and Health Promotion and the Department of Biological and Clinical Psychology of the University of Jena through a telephone survey. The local ethics committee of the University of Jena approved the study protocol; the study was performed in accordance with the Helsinki Declaration. All volunteers provided written informed consent prior to participation.

Study design

A repeated-measures design was performed. Each subject participated in two QST sessions (measurement session 1: T1; and measurement session 2: T2) separated by 10.0±2.9 weeks (mean ± SD). The whole QST test battery was conducted on two anatomical locations: on the lower back (paraspinal lumbar area) and the dorsum of the dominant hand. The area of measurement for the lower back was between vertebrae L2 and L5 with a mean distance of 4.4±1.8 cm (mean ± SD) lateral from the spinous process. The two areas measured (lower back and hand dorsum) correspond with the contact area of the thermode (3×3 cm). The size of the area was

Table 1 Descriptive characteristics of all participants (N=22)

Characteristics	Healthy participants		
	Total	Male	Female
Number of participants, N (%)	22 (100)	10 (45)	12 (55)
Age, years, mean ± SD (range)	46.6±13.0 (23–61)	38.2±13.1 (23–56)	54.2±6.8 (37–61)
Age distribution			
20–29 years, N	4	4	0
30–39 years, N	2	1	1
40–49 years, N	3	2	1
50–59 years, N	11	3	8
60–69 years, N	2	0	2

Abbreviation: N, number of participants.

marked using a skin marker. All participants were right-handed. The test instructions were given by reading the standard QST instructions for all subjects. The measurements were performed in a quiet room (room temperature: 23.0°C±2.2°C, mean ± SD). To control for the influence of skin temperature, the skin temperatures of the lower back and hand dorsum were recorded before administering the QST. All data were collected by one examiner (HN), trained by the DFNS. The laboratory is certified by the DFNS (registration number: 36180814).

QST protocol

QST was performed according to the standardized protocol of the DFNS.[2,10]

Thermal detection and pain thresholds as well as the number of paradoxical heat sensations (PHS)

Thermal testing was conducted using the thermal stimulator Thermal Sensory Analyzer II (TSA; Medoc, Ramat Yishai, Israel) with a contact area of the thermode equaling 9 cm^2. Cold detection threshold (CDT), warm detection threshold (WDT), number of PHS using the thermal sensory limen (TSL) procedure of alternating cold and warm stimuli, cold pain threshold (CPT), and heat pain threshold (HPT) were each assessed using the standard protocol of DFNS. The baseline temperature was 32°C, with a lower cutoff temperature at 0°C and upper cutoff temperature at 50°C, and the ramp rate for all thermal stimuli was 1°C/s.[2,10]

Mechanical detection threshold (MDT)

MDT for touch was assessed by using a standardized set of modified von Frey filaments (diameter 0.5 mm, Optihair$_2$-Set Marstock Nervtest, Schriesheim, Germany), which exert forces between 0.25 mN and 512 mN (factor two progression). Using the "method of limits", the final threshold was defined as the geometric mean of five series of ascending and descending stimulus intensities.[32]

Mechanical pain threshold (MPT)

MPT was measured using a set of standard pinprick stimulators (cylindrical tip, 250 µm tip diameter) with fixed stimulus intensities that exerted forces of 8, 16, 32, 64, 128, 256, and 512 mN (MRC Systems GmbH, Heidelberg, Germany). The stimulators were applied in ascending order until the first perception of sharpness was detected. MPT was determined using the "method of limits". The final threshold was the geometric mean of five series of ascending and descending stimuli intensities.

Mechanical pain sensitivity (MPS) and dynamic mechanical allodynia (DMA)

Pain induced by punctate mechanical stimuli was measured using the same standard pinprick stimulators as for MPT. To obtain MPS for pinprick-evoked pain, all seven pinprick stimuli were applied in balanced order, five times each stimulus. MPS was defined as the geometric mean of the given stimuli as in the standard protocol. To avoid effects of sensitization or fatigue, the successive stimuli were not applied at the same spot of skin, but some millimeters away from the previously stimulated spot. Following each stimulus, participants were asked to rate the experienced pain intensity for each stimulus on an NRS, with zero indicating "no pain" and 100 indicating "maximal imaginable pain". Pain to light touch (DMA) was assessed by light stroking with a cotton wisp (3 mN), a Q-tip fixed to an elastic strip (100 mN), and a soft makeup brush (200–400 mN). The set of the three light tactile stimulators were intermingled with the pinprick stimuli in balanced order. If the stroking stimuli were perceived as painful, participants were asked to give a rating for the amount of perceived pain using the same NRS (0–100).

Wind-up ratio (WUR)

The perceptual correlate of temporal pain summation to repetitive pinprick stimuli (WUR) was assessed by a series of 10 pinprick stimuli (256 mN) with 1 Hz repetition rates. The participants were asked to give a pain rating representing the pain at the end of the stimuli series using the 0–100 NRS. The pain ratings to single pinprick stimulation were compared with those of 10 repeated stimuli. To determine the WUR, the ratio of the mean pain rating of the series divided by the mean pain rating of a single stimulus was calculated after five trials.

Vibration detection threshold (VDT)

VDT was measured using a standardized Rydel–Seiffer graded tuning fork (64 Hz, 8/8 scale) that was placed over a bony prominence (back: spinous processes of the vertebrae between L2 and L5, hand: ulnar styloid process) according to the protocol of DFNS.[2,10,11] Volunteers reported the disappearance of the vibration. VDT was determined as the average of three consecutive measurements of the amount of time to disappearance.

Pressure pain threshold (PPT)

PPT was assessed over a muscle on the test areas (back: lumbar paraspinal, hand: pollicis muscles/thenar eminence) using a pressure gauge device (FDN200, Wagner Instruments,

Greenwich, CT, USA) with a probe area of 1 cm^2 and that exerts pressure up to 2000 kPa. The PPT was determined with three series of ascending stimulus intensities, each applied at an increasing ramp of 50 kPa/s.

Data analysis

QST data analysis was performed as recommended and described previously.[2,10,12] The mean thresholds for each subject were calculated using Microsoft Excel (Microsoft Office 2013; Microsoft Corporation, Redmond, WA, USA). These results were summarized in a single-sheet QST report form for each subject and were used for further statistical analysis.

All QST parameters (except CPT, HPT, and VDT) were logarithmically transformed (base 10) to achieve a (secondary) normal distribution.[10] A small constant (0.1) was added to pain ratings for pinprick (MPS) prior to log-transformation to avoid a loss of values due to zero rating.[33] It has been already shown that QST parameters were normally (or log-normally) distributed in a healthy population.[2]

Analysis of reliability and agreement

The investigation of how similar two tests are is complex and reaches beyond the commonly used test–retest reliability.[34] Following recommendations[34] and previous methodology,[27] we investigated the following: differences between measurements T1 and T2, correlations, intraclass correlation coefficients (ICCs), and Bland–Altman plots.

- Differences between measurements (T1 – T2): Quantitative differences between both QST sessions (T1 and T2) were investigated with paired samples *t*-test.
- Correlations between T1 and T2: Pearson's product–moment correlation coefficient was used to determine the strength of correlation between T1 and T2. Correlation coefficients were interpreted as follows: $|r| \leq 0.29$ negligible, 0.30–0.49 low, 0.50–0.69 moderate, 0.70–0.89 high, and >0.90 very high.[35] Additionally, Tables S3 and S4 report the Pearson's product–moment correlation coefficient for normally distributed QST parameters and Spearman's rank correlation coefficient for nonnormally distributed parameters.
- IC coefficient (ICC): ICC was calculated using the two-way random-effects analysis of variance (ANOVA) model, type absolute agreement. The ICC values were interpreted as follows: ICC ≤0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and >0.80 almost perfect.[36]
- Bland–Altman plots: Bland–Altman plots were used in order to assess the level of agreement between T1 and T2.[31,37] These plots give a transparent visual presentation of potential bias of the data and the so-called LoAs.[34] The plots display differences between both QST measurements (T1 – T2) against the mean values of both QST measurements (T1 + T2)/2) for each subject. These plots represent the average bias: if differences are systematically greater than zero, then the QST measurement in T1 is systematically higher than the QST measurement in T2; conversely, if differences are systematically less than zero, then the measurement in T1 is systematically lower than that in T2. The LoAs (mean differences ±1.96× SD) of the Bland–Altman plots were used as a central outcome for the determination of the agreement between the two QST assessments. These LoAs can be interpreted in the following way. If the study were to be repeated, the difference between measurements should lie within these limits in 95% of all cases. Conclusions about reliability should be drawn from the mean of differences (average bias) as well as the LoAs: if the differences are small (ie, the LoAs are small) and the mean of the differences is near zero, the test can be considered reliable.

In addition, the standard error of measurement (SEM) and the minimum detectable difference (MDD) were calculated. SEM was determined as the square root of the mean square error term from the repeated-measures ANOVA. The MDD was calculated using the following formula: MDD (SEM $\times 1.96 \times 2^{1/2}$).[30,38]

Confirming our findings of our primary statistical analysis, we performed an additional statistical analysis without outliers of each parameter (Tables S5 and S6, Figures S1 and S2). The outliers were defined as the values that were more than 1.5 × interquartile range (IQR) beyond the 25th and 75th percentiles.

All statistical calculations were performed using SPSS Statistics 22 (IBM Corporation, Armonk, NY, USA). We considered the analysis of each parameter as a separate hypothesis, and therefore, no adjustment for multiple comparisons was needed. Hence, the significance level was set to 0.05 for each statistical test. For the Bland–Altman plots, the software package R (version 3.2.4; R Core Team, Vienna, Austria) was used.

Results

All participants completed the study. Tables 2 and 3 show the results for the QST variables (log-transformed data for CDT, WDT, TSL, PPT, MPT, MPS, WUR, MDT, and raw data for CPT, HPT, and VDT) of the lower back and dominant hand for the total sample size (N=22). The results present

Table 2 Statistical analysis of QST values for the lower back in healthy volunteers (N=22)

Parameter	Difference (T1 – T2)		Correlation		ICC			LoAs	SEM	MDD
	Mean ± SD (95% CI of mean)	*p*	*r*	*p*	ICC	*p*	(95% CI)	Lower LoA – upper LoA		
Back										
CDT_{log} (Δ*T*), °C	0.17 ± 0.27 (0.05–0.30)	0.007*	0.18	0.432	0.14	0.212	(–0.18 to 0.47)	–0.36 to 0.71	0.19	0.53
WDT_{log} (Δ*T*), °C	0.07 ± 0.26 (–0.05 to 0.18)	0.250	–0.53	0.012*	–0.51	0.995	(–0.79 to 0.10)	–0.45 to 0.58	0.19	0.52
TSL_{log} (Δ*T*), °C	0.08 ± 0.17 (0.01–0.16)	0.038*	0.26	0.248	0.23	0.119	(–0.14 to 0.57)	–0.26 to 0.42	0.12	0.34
CPT, °C	–6.35 ± 10.34 (–10.94 to –1.77)	0.009*	0.35	0.109	0.26	0.070	(–0.10 to 0.58)	–26.62 to 13.91	7.31	20.26
HPT, °C	–0.14 ± 1.69 (–0.89 to 0.61)	0.708	0.83	<0.001**	0.80	<0.001**	(0.58–0.91)	–3.45 to 3.18	1.20	3.31
PPT_{log}, kPa	0.02 ± 0.17 (–0.05 to 0.10)	0.492	0.64	0.001**	0.65	<0.001**	(0.32–0.84)	–0.30 to 0.35	0.12	0.33
MPT_{log}, mN	0.13 ± 0.35 (–0.02 to 0.28)	0.099	0.52	0.013*	0.50	0.005*	(0.13–0.75)	–0.55 to 0.81	0.24	0.68
MPS_{log}, NRS	–0.12 ± 0.25 (–0.23 to 0.01)	0.035*	0.86	<0.001**	0.84	<0.001**	(0.62–0.93)	–0.60 to 0.36	0.17	0.48
WUR_{log}, ratio	0.02 ± 0.23 (–0.09 to 0.12)	0.752	0.51	0.016*	0.52	0.007*	(0.13–0.77)	–0.43 to 0.46	0.16	0.45
MDT_{log}, mN	0.06 ± 0.40 (–0.12 to 0.23)	0.510	0.22	0.329	0.22	0.159	(–0.22 to 0.58)	–0.73 to 0.84	0.28	0.78
VDT, *x*/8	0.06 ± 0.85 (–0.31 to 0.44)	0.739	0.65	0.001**	0.62	0.001**	(0.27–0.82)	–1.60 to 1.72	0.60	1.66

Notes: QST parameters were logarithmically transformed (except for CPT, HPT, and VDT), according to recommendations of Rolke et al.[2] $Index_{log}$ denotes QST parameters for which calculations are based on log-transformed data. T1, measurement session 1; T2, measurement session 2; Δ*T*, difference in temperature to the 32°C baseline; level of significance: *$p \leq 0.05$; **$p \leq 0.001$.

Abbreviations: QST, quantitative sensory testing; N, number of participants; CDT, cold detection threshold; WDT, warm detection threshold; TSL, thermal sensory limen; CPT, cold pain threshold; HPT, heat pain threshold; PPT, pressure pain threshold; MPT, mechanical pain threshold; MPS, mechanical pain sensitivity; WUR, wind-up ratio; MDT, mechanical detection threshold; VDT, vibration detection threshold; NRS, numerical rating scale; ICC, intraclass correlation coefficient; LoA, limits of agreement according to Bland and Altman;[31] MDD, minimum detectable difference.

Table 3 Statistical analysis of QST values for the dominant hand in healthy volunteers (N=22)

Parameter	Difference (T1 – T2)		Correlation		ICC			LoAs	SEM	MDD
	Mean ± SD (95% CI of mean)	*p*	*r*	*p*	ICC	*p*	(95% CI)	Lower LoA – upper LoA		
Hand										
CDT_{log} (Δ*T*), °C	0.07 ± 0.30 (–0.06 to 0.20)	0.279	0.13	0.560	0.11	0.300	(–0.31 to 0.50)	–0.51 to 0.65	0.21	0.57
WDT_{log} (Δ*T*), °C	0.05 ± 0.25 (–0.06 to 0.16)	0.347	0.60	0.003*	0.60	0.001**	(0.25–0.81)	–0.43 to 0.53	0.17	0.48
TSL_{log} (Δ*T*), °C	0.06 ± 0.23 (–0.04 to 0.16)	0.243	0.50	0.017*	0.50	0.007*	(0.12–0.75)	–0.39 to 0.51	0.16	0.45
CPT, °C	–1.51 ± 8.52 (–5.29 to 2.27)	0.414	0.39	0.071	0.39	0.033*	(–0.02 to 0.69)	–18.22 to 15.20	6.03	16.71
HPT, °C	–0.12 ± 1.79 (–0.91 to 0.67)	0.752	0.80	<0.001**	0.81	<0.001**	(0.59–0.92)	–3.63 to 3.38	1.26	3.50
PPT_{log}, kPa	0.03 ± 0.15 (–0.04 to 0.10)	0.341	0.52	0.014*	0.49	0.008*	(0.11–0.75)	–0.27 to 0.33	0.10	0.29
MPT_{log}, mN	0.23 ± 0.41 (0.04–0.41)	0.017*	0.48	0.023*	0.42	0.011*	(0.03–0.70)	–0.58 to 1.03	0.29	0.80
MPS_{log}, NRS	–0.10 ± 0.41 (–0.28 to 0.08)	0.265	0.66	0.001**	0.66	<0.001**	(0.34–0.84)	–0.90 to 0.70	0.29	0.80
WUR_{log}, ratio	–0.01 ± 0.26 (–0.13 to 0.10)	0.834	0.29	0.199	0.27	0.112	(–0.18 to 0.62)	–0.53 to 0.51	0.18	0.51
MDT_{log}, mN	0.09 ± 0.38 (–0.08 to 0.25)	0.304	0.32	0.153	0.31	0.072	(–0.11 to 0.64)	–0.66 to 0.83	0.27	0.75
VDT, *x*/8	0.07 ± 0.58 (–0.19 to 0.33)	0.590	0.63	0.002*	0.62	0.001**	(0.28–0.82)	–1.08 to 1.21	0.41	1.14

Notes: QST parameters were logarithmically transformed (except for CPT, HPT, and VDT), according to recommendations of Rolke et al.[2] $Index_{log}$ denotes QST parameters for which calculations are based on log-transformed data. T1, measurement session 1; T2, measurement session 2; Δ*T*, difference in temperature to the 32°C baseline; level of significance: *$p \leq 0.05$; **$p \leq 0.001$.

Abbreviations: QST, quantitative sensory testing; N, number of participants; CDT, cold detection threshold; WDT, warm detection threshold; TSL, thermal sensory limen; CPT, cold pain threshold; HPT, heat pain threshold; PPT, pressure pain threshold; MPT, mechanical pain threshold; MPS, mechanical pain sensitivity; WUR, wind-up ratio; MDT, mechanical detection threshold; VDT, vibration detection threshold; NRS, numerical rating scale; ICC, intraclass correlation coefficient; LoA, limits of agreement according to Bland and Altman;[31] MDD, minimum detectable difference.

the differences between the measurements T1 and T2, their mean, SD, and 95% CI of mean. In addition, the results of the statistical analysis of differences between both measurements (T1 and T2), correlations, ICCs, LoAs for the Bland–Altman plot, SEMs, and MDDs are reported (Tables 2 and 3).

Additionally, we analyzed our data concerning outliers for each test parameter and for each test area (back and hand). Tables S5 and S6 outline the statistical analysis without outliers for all calculations (differences in terms of T1 – T2, correlations, ICCs, LoAs, SEM, and MDD). Furthermore, Tables S3 and S4 outline the correlation analysis by Pearson's product–moment correlation (normal distribution) and the Spearman's rank correlation (nonnormal distribution).

Differences between measurements (T1 – T2)

Back

The comparison of the means of T1 and T2 showed significant differences ($p \leq 0.05$) for CDT, TSL, CPT, and MPS (Table 2).

Hand

On the dominant hand dorsum, significant differences were observed for MPT only (Table 3).

Correlations between T1 and T2

Back

The correlation analysis revealed significant high correlations ($r \geq 0.70$; $p \leq 0.001$) for HPT and MPS, as well as moderate correlations ($r \geq 0.50$; $p \leq 0.013$) for WDT, PPT, MPT, WUR, and VDT (Table 2).

Hand

A significant high correlation between T1 and T2 was shown for HPT ($r=0.80$; $p<0.001$; Table 3). Moderate correlations were found for WDT, TSL, PPT, MPS, and VDT (ranging from $r=0.50$ to $r=0.66$; all $p<0.017$). Lower, but still significant, correlation was found for MPT ($r=0.48$; $p<0.023$). Tables S3 and S4 outline the correlation analysis of QST on the lower back and on the dominant hand in healthy volunteers (N=22) according to the normal distribution.

ICC analysis

Back

The ICC analysis demonstrated an almost-perfect ICC for MPS (ICC: 0.84; $p<0.001$) at the lower back (Table 2). HPT, PPT, and VDT achieved substantial ICCs (ranging from ICC 0.62 to ICC 0.80; $p \leq 0.001$), and moderate ICCs were observed for MPT and WUR (ICC 0.50 and ICC 0.52; $p \leq 0.011$).

Hand

An almost-perfect ICC was shown for HPT (ICC 0.81; $p<0.001$), while the ICCs for MPS and VDT (ICC 0.66 and ICC 0.62; $p \leq 0.001$) were substantial. Moderate ICCs were observed for WDT, TSL, PPT, and MPT (ranging from ICC 0.42 to ICC 0.60; $p \leq 0.011$) on the hand dorsum (Table 3). Fair ICC was found for CPT (ICC =0.39; $p \leq 0.033$).

Bland–Altman plots

Figures 1 and 2 depict the Bland–Altman plots for all QST parameters. The LoAs of all parameters contained the number zero for all variables on lower back and hand dorsum.

Back

The LoAs varied substantially between QST parameters, being lowest for WUR (–0.27 to 0.35) and highest for the thermal pain thresholds (CPT: –26.62 to 13.91; HPT: –3.45 to 3.18).

Hand

The LoAs varied between the QST parameters, being the lowest for PPT (–0.27 to 0.33) and highest for the thermal pain threshold (CPT: –18.22 to 15.20; HPT: –3.63 to 3.38).

All QST values in original units (raw data) with mean, SD, 95% CI of mean, median, 25th and 75th percentiles for T1, T2, the differences of both measurements (T1 and T2), the SEM, and the MDD are presented in Tables S1 and S2 for the back and the hand, respectively.

SEM results

SEM and MDD are given in Tables 2 and 3 as complements to the reliability and agreement measures. These measures are important as they determine the clinically significant measures for each of the QST parameters. As can be seen in Tables 2 and 3, the SEM values were lower for the hand compared to those for the back for WDT, CPT, PPT, MDT, and VDT, indicating that these QST parameters are precise at the hand over a 10-week period. In contrast, the SEM values for CDT, TSL, HPT, MPT, MPS, and WUR are lower at the back compared to the hand, indicating that these parameters are more precise at the back than at the hand over the 10-week period.

Discussion

The primary aim of this study was to undertake a detailed analysis of reliability and agreement, with a time interval of 10 weeks, for the standardized QST procedure according to the German Research Network on Neuropathic Pain (DFNS). To our knowledge, the present study is the first indicating the normal variations for each parameter of the broadly used QST protocol of the DFNS[2,8,10–12,14,39,40] over a standard intervention time (10 weeks). This helps to define ranges of normal variation to provide a basis for using QST as a tool in 10-week interventional trials and to assess interventional outcomes on an individual basis.

ICC values

High ICC values were found for some QST parameters, such as HPT, at the back. This is in accordance with previous reports.[20,30] However, these findings were not confirmed in all of the available studies (eg, only fair reliability for CPT[17,22]).

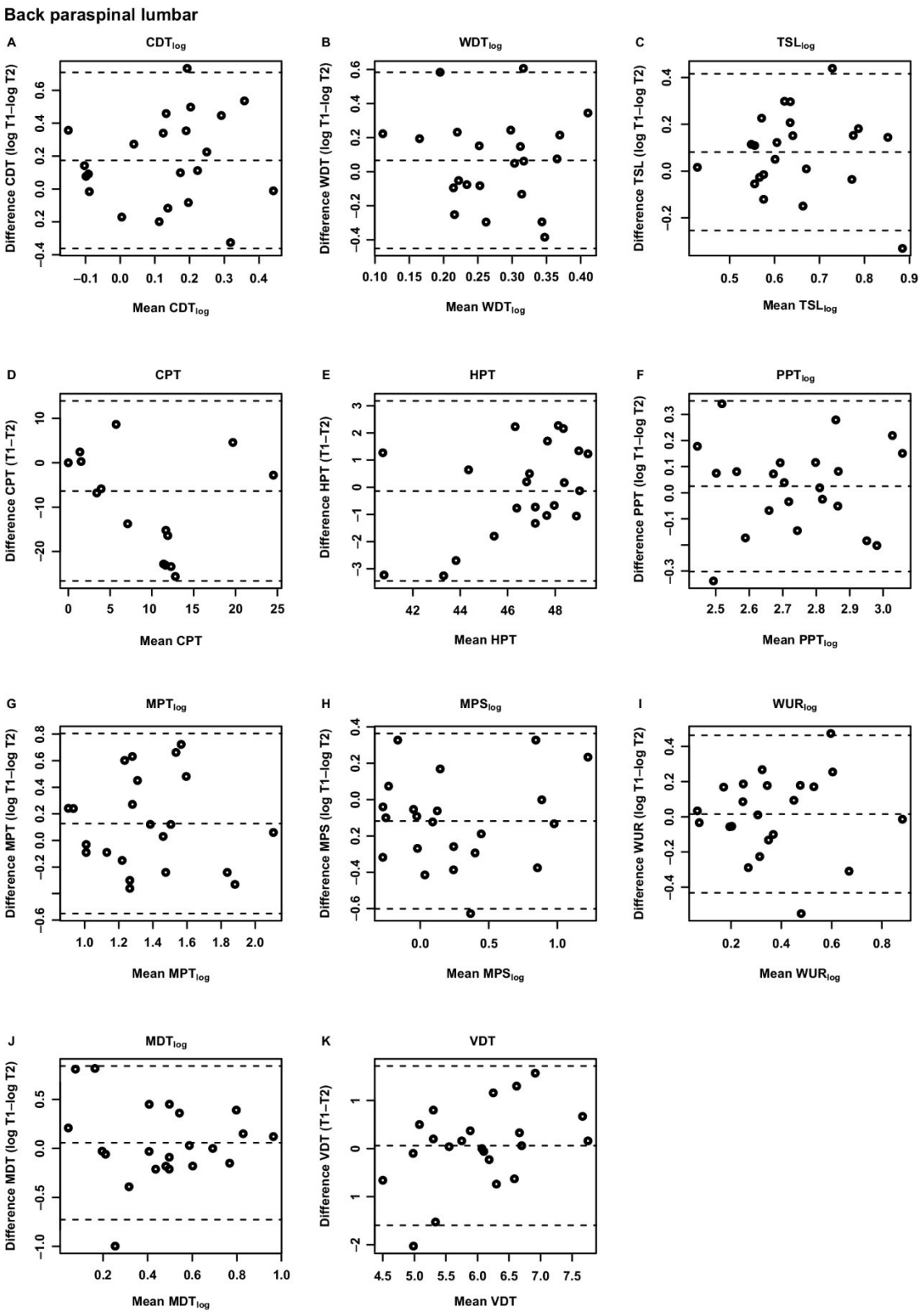


Figure 1 Bland–Altman plots of the QST parameters for the lower back of healthy volunteers (N=22).

Notes: (**A**) CDT_{log}; (**B**) WDT_{log}; (**C**) TSL_{log}; (**D**) CPT; (**E**) HPT; (**F**) PPT_{log}; (**G**) MPT_{log}; (**H**) MPS_{log}; (**I**) WUR_{log}; (**J**) MDT_{log}; (**K**) VDT; T1, measurement session 1; T2, measurement session 2. Bland–Altman plots with the differences between T1 and T2 values (vertical axis) plotted against the mean of each T1 and T2 value (horizontal axis) of each participant. The middle horizontal dashed line represents the mean difference between T1 and T2 of all subjects; upper and lower dashed lines indicate the limits of agreement (upper and lower limits of agreement, mean difference ±1.96× SD).

Abbreviations: CDT, cold detection threshold; WDT, warm detection threshold; TSL, thermal sensory limen; CPT, cold pain threshold; HPT, heat pain threshold; PPT, pressure pain threshold; MPT, mechanical pain threshold; MPS, mechanical pain sensitivity; WUR, wind-up ratio; MDT, mechanical detection threshold; VDT, vibration detection threshold.

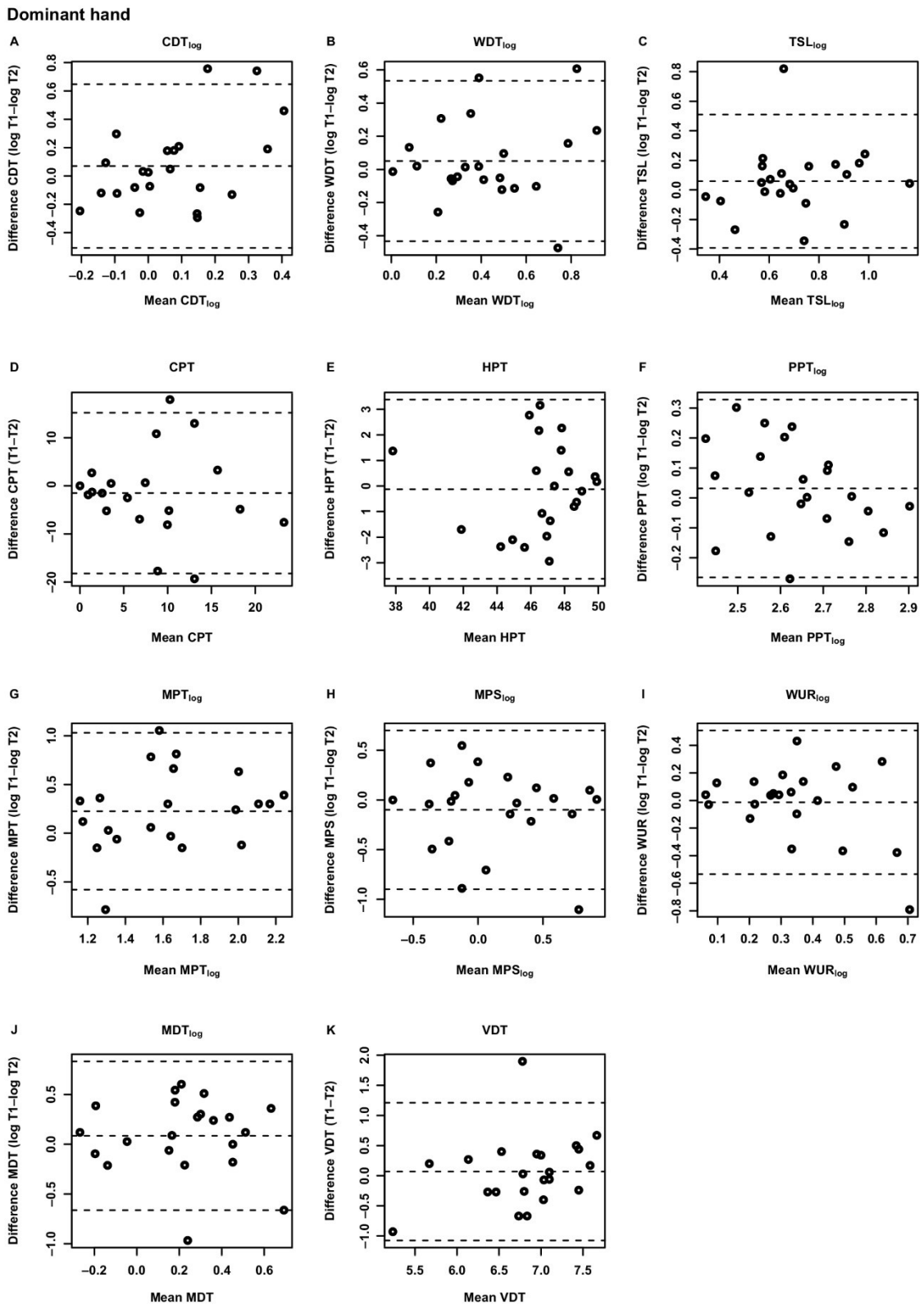


Figure 2 Bland–Altman plots of the QST parameter for the dominant hand of healthy volunteers (N=22).

Notes: (**A**) CDT$_{log}$; (**B**) WDT$_{log}$; (**C**) TSL$_{log}$; (**D**) CPT; (**E**) HPT; (**F**) PPT$_{log}$; (**G**) MPT$_{log}$; (**H**) MPS$_{log}$; (**I**) WUR$_{log}$; (**J**) MDT$_{log}$; (**K**) VDT; T1, measurement session 1; T2, measurement session 2. Bland–Altman plots with the differences between T1 and T2 values (vertical axis) plotted against the mean of each T1 and T2 value (horizontal axis) of each participant. The middle horizontal dashed line represents the mean difference between T1 and T2 of all subjects; upper and lower dashed lines indicate the limits of agreement (upper and lower limits of agreement, mean difference ±1.96 × SD).

Abbreviations: CDT, cold detection threshold; WDT, warm detection threshold; TSL, thermal sensory limen; CPT, cold pain threshold; HPT, heat pain threshold; PPT, pressure pain threshold; MPT, mechanical pain threshold; MPS, mechanical pain sensitivity; WUR, wind-up ratio; MDT, mechanical detection threshold; VDT, vibration detection threshold.

Wasner and Brock[41] investigated three different measuring points and ascertained a better ICC for CPT for an interval of 1 day versus an interval of 21 days. Those results are in line with our findings on the CPT, showing lower reliabilities for a time interval of 10 weeks.

Moderate-to-substantial ICCs for the PPT on the hand and back over 10 weeks suggest that the mechanical QST parameters are stable over time. Previous studies reported good-to-excellent ICCs over short-term intervals for the PPT on the face, neck, forearm, finger, and knee.[16,25,28,42] Andersen et al,[43] who investigated QST changes before and 1 year after breast cancer surgery, describe a good reliability for MPT determined 1 week apart. One study revealed substantial reliability for VDT on the back and hand, as well as moderate reliability for MPS and slight reliability for WUR on the face.[28] Felix and Widerstrom-Noga[17] even showed an excellent ICC for VDT within a short-term interval, while good-to-excellent ICCs were reported for MDT.[17,25] Overall, our results are in line with lower ICCs for the mechanical QST parameters, when tested over short-term intervals.[38]

In conclusion, test–retest reliability is sufficient for most of the QST measures.

Agreement

Our data show smaller as well as larger LoAs for different QST parameters. LoAs have been investigated only rarely with regard to QST parameters. Hirschfeld et al[27] reported LoAs for repeated QST measures for the hands at the same day in a longitudinal investigation of developmental changes in somatosensation in children. Another study investigated some QST parameters in the course of breast cancer surgery, reporting the LoAs for forearm, leg, and breast.[43] The reported LoA (Table 5 in the study by Andersen et al[43]) are larger or in the same range as our data. This is surprising as the time difference between QST examinations in our study was larger. Possible reasons for the narrower limits in our study might be due to the population examined (patients in some studies vs healthy subjects in our study) and the highly standardized procedures used in our laboratory (standardized QST protocol, standardized test instructions, meeting the extensive requirements for certification of our QST laboratory).[44] While ICCs are often nonsignificant, LoAs exhibit rather favorably small and consistent 95% CIs, which appear to be counterintuitive. Bland–Altman plots (Figures 1 and 2) reveal that many of the individual QST values are located within a small interval, suggesting the existence of a mean value with a small, but random variation (eg, CPT on the lower back; Figure 1D). This lack of systematic variation alone is sufficient to explain the lack of a significant correlation. However, it is exactly this small data range of some of the QST parameters, particularly in healthy subjects, which makes QST a suitable measure to detect both deviations from the norm, as may be seen in patients,[8,39,40] as well as changes induced by treatments. The question of how stable a measure is can be answered by analyzing how similar two measurements are, and this question is best answered by using the LoAs.[45,46]

SEM results

SEM as an absolute reliability index that reflects the agreement between repeated measures within each individual is suitable in interventional trials to calculate the MDD values for determining the sample sizes.[30] To our knowledge, there is only one study[30] providing SEMs for time intervals longer than 1 week for four QST parameters, namely, CPT, HPT, PPT, and WUR. Our SEM values are mainly in line with the results of Marcuzzi et al.[30] Thus, our data confirm the lower SEM values at the hand for CPT and PPT, indicating that CPT and PPT are more precise when conducted at the hand than at the back. Accordingly, HPT showed lower SEM values at the back in comparison to the hand, both in our study and in the study by Marcuzzi et al.[30] In contrast to that study,[30] SEM values for WUR were higher at the hand in comparison to those of the back, indicating that WUR in our study is more precise at the back than at the hand. This difference cannot be explained by outliers (Tables S5 and S6).

As we used the comprehensive QST protocol of the DFNS network, we provide, for the first time, SEMs for all QST parameters expected. Our data indicate that SEM values for CDT, TSL, MPT, and MPS were lower at the back compared to those at the hand, indicating that these QST parameters are precise measures when conducted at the back over a 10-week period. In contrast, SEM values for WDT, MDT, and VDT were lower at the hand compared to the back for this period.

Future directions and limitations

Knowledge of the variability of QST parameters is a prerequisite to assess meaningful changes of any kind of intervention on these parameters (which should be larger than the variability). One strength of our study is the application of the comprehensive QST protocol, including detection thresholds (CDT, WDT, TSL, MDT, and VDT), as well as pain thresholds and related pain parameters (HPT, PPT, MPT, MPS, and WUR), over a period of 10 weeks. Our data indicate that QST parameters are suitable for individual monitoring over 10 weeks. Furthermore, our study provides a set

of useful items to depict reliability and agreement. Beyond the estimation of effects from interventional trials, our data might contribute to the expression of a treatment effect on QST with regard to the LoAs. The data for each subtest may serve as the norms in order to evaluate the deviation of a patient's value. The effects of an intervention would then be expressed as the possible change in the deviation, or, ideally, if the subject's values lie within the LoAs after intervention, as normalization. The use of the LoAs would overcome some of the issues of all correlation-based measures, such as test–retest reliability, namely, that the absolute values may well decrease or increase over time, but as long as the ranking of the individual within the cohort is similar between the two measurements, the correlation is high. With regard to the estimation of a treatment effect, this phenomenon may represent a challenge. One example of a QST subtest where this could be relevant is the PPT, which has been shown to change from the first to the second measurement in particular, while still exhibiting high test–retest reliabilities.[30,42] Even though we did not find such dramatic differences between T1 and T2, a similar trend was seen in our data. Here, the LoAs may be more suitable to detect an intervention-induced effect.

A DFNS-trained researcher (HN) performed the QST assessment in all subjects at both T1 and T2. Our approach was similar to a recent study on long-term reliability on some QST parameters,[30] where one investigator performed the QST measurements. Since the aim of our study was to determine the stability (test–retest reliability and agreement) over longer periods under ideal conditions, we chose 1) a highly standardized protocol, namely, the DFNS protocol, 2) healthy, pain-free volunteers, and 3) one, highly trained investigator. The rationale for this approach was the assumption that, if QST did not show sufficient long-term reliability and agreement under these highly standardized conditions, further investigation of this research question would be obsolete as QST would not be a suitable measure for interventional trials on pain. Moreover, a highly standardized approach represents also the method of choice within a clinical trial. Nonetheless, it can be questioned whether such an approach achieves a sufficient degree of ecological validity for clinical practice, where often several clinicians perform the tests. However, this question was not the focus of the present study. It would be of interest to assess the interrater long-term reliability, similarly to short-term reliability.[29]

Generally, our results as well as the data from Marcuzzi et al[30] do not necessarily reflect the results from literature discussed herein. Both studies investigate healthy volunteers, while the results on reliability and agreement in most of the studies are derived from patient populations, or, as in the case of the data from Hirschfeld et al,[27] from somewhat atypical populations such as children. Thus, generalizability of our results to a patient population remains to be confirmed. Similar to the data of Marcuzzi et al,[30] some of the QST parameters seem to show considerable adjustments over time and it remains unclear whether this is an effect related to learning. Since Marcuzzi et al[30] found the strongest differences between T1 and T2, it may be recommended to assess two baseline measurements in a clinical trial setting, in order to avoid a confounding effect within the treatment phase of the trial.

Even though our results are to a large extent in line with a recent published study with a larger sample,[30] one relevant limitation of this study is its small sample size (N=22). Nonetheless, assuming a power of 80%, large effects with Cohen's $d>0.6$ could be sufficiently detected by our study. Our study asks whether QST is suitable for monitoring individuals over 10 weeks, which might be useful for monitoring the effects of interventions in clinical trials on pain. Within a clinical trial, the most relevant question is whether the difference found between two interventions is clinically meaningful. This is usually measured with a visual analogue scale or other markers as main outcomes, directly related to the patients' experience of pain. A clinically meaningful effect may well represent a 30%–50% change of the main outcome with regard to baseline,[47] which can be considered a strong effect. The power of a clinical trial may be increased by investigating a large number of patients, so that smaller changes also of the main outcomes, eg, a 10% change from baseline, become statistically significant. Such a trial would be able to detect small effect sizes. However, the results, even though statistically significant, would clinically not be relevant. In conclusion, within a clinical trial setting, the achieved power may well be sufficient to detect a clinically significant change. However, further studies, in particular within relevant patient groups, are urgently needed.

There are some technological limitations. These limitations are inherent to the equipment and not related to the particular protocol used. The CPT normally has a cutoff temperature of 0°C. Some participants do not push the button before the thermode stops the cooling automatically, which means no painful cold experience is reported. Frey filaments are limited in their range from 0.25 mN to 512 mN.[2,10] Some healthy participants are able to detect the thinnest hair with a force of 0.25 mN. These limitations may induce a ceiling effect.

To date, test–retest reliability and agreement of QST are inconclusive. Inconsistent results ranging from poor

reliability to excellent reliability are related to lack of standardization of the procedures, including the use of different equipments, algorithms, populations, and statistical methods.[5] Moreover, studies are often selective in their use of tests, investigating the reliability of only thermal QST modalities[20,22,41] or of some selected parameters of the QST.[17,25,43] The time interval is also rather inconsistent and ranges from 3 days to 3 weeks.[17,20,22,25,43] To avoid the standardization issues, we followed the standard QST protocol according to the German Research Network on Neuropathic Pain (DFNS)[2,10] and used an interval similar to that often used in clinical trials on pain.

A major problem is that the synonymous use of the terms reliability and agreement is inappropriate.[48] The different statistical approaches make a comparative interpretation of results almost impossible. A critical review on test–retest studies in QST identified considerable heterogeneity in statistical evaluations and recommends minimal methodological requirements and a protocol for reporting test–retest data.[34]

Conclusion

This study presents test–retest data, SEMs, MDDs, and LoAs for two highly standardized QST measurements for the comprehensive QST protocol over a period of 10 weeks. With regard to the additionally tested retest reliability, the results are generally in line with recently published data.[30] Our data constitute the first step to define LoAs for each subtest of the standardized QST. Such data are a prerequisite when QST should serve as a basis for interventional trials as an outcome, which might be relevant for clinical research.

Acknowledgments

Jan Vollert is acknowledged for his help in statistical analysis of the QST data. We thank all the healthy volunteers for their participation in our study.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Pavlakovic G, Petzke F. The role of quantitative sensory testing in the evaluation of musculoskeletal pain conditions. *Curr Rheumatol Rep*. 2010;12(6):455–461.
2. 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):231–243.
3. Vollert J, Attal N, Baron R, et al. Quantitative sensory testing using DFNS protocol in Europe: an evaluation of heterogeneity across multiple centers in patients with peripheral neuropathic pain and healthy subjects. *Pain*. 2016;157(3):750–758.
4. Backonja MM, Attal N, Baron R, et al. Value of quantitative sensory testing in neurological and pain disorders: NeuPSIG consensus. *Pain*. 2013;154(9):1807–1819.
5. Chong PS, Cros DP. Technology literature review: quantitative sensory testing. *Muscle Nerve*. 2004;29(5):734–747.
6. Cruccu G, Sommer C, Anand P, et al. EFNS guidelines on neuropathic pain assessment: revised 2009. *Eur J Neurol*. 2010;17(8):1010–1018.
7. Shy ME, Frohman EM, So YT, et al; Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. Quantitative sensory testing: report of the therapeutics and technology assessment subcommittee of the American Academy of Neurology. *Neurology*. 2003;60(6):898–904.
8. Puta C, Schulz B, Schoeler S, et al. Somatosensory abnormalities for painful and innocuous stimuli at the back and at a site distinct from the region of pain in chronic back pain patients. *PLoS One*. 2013;8(3):e58885.
9. Spohn D, Musial F, Rolke R. Naturopathic reflex therapies for the treatment of chronic pain – part 2: quantitative sensory testing as a translational tool. *Forsch Komplementmed*. 2013;20(3):225–230.
10. 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.
11. Magerl W, Krumova EK, Baron R, Tolle T, Treede RD, Maier C. Reference data for quantitative sensory testing (QST): refined stratification for age and a novel method for statistical comparison of group data. *Pain*. 2010;151(3):598–605.
12. Pfau DB, Krumova EK, Treede RD, et al. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): reference data for the trunk and application in patients with chronic postherpetic neuralgia. *Pain*. 2014;155(5):1002–1015.
13. Vollert J, Mainka T, Baron R, et al. Quality assurance for quantitative sensory testing laboratories: development and validation of an automated evaluation tool for the analysis of declared healthy samples. *Pain*. 2015;156(12):2423–2430.
14. Maier C, Baron R, Tölle TR, et al. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): somatosensory abnormalities in 1236 patients with different neuropathic pain syndromes. *Pain*. 2010;150(3):439–450.
15. Agostinho CM, Scherens A, Richter H, et al. Habituation and short-term repeatability of thermal testing in healthy human subjects and patients with chronic non-neuropathic pain. *Eur J Pain*. 2009;13(8):779–785.
16. Cathcart S, Pritchard D. Reliability of pain threshold measurement in young adults. *J Headache Pain*. 2006;7(1):21–26.
17. Felix ER, Widerstrom-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.
18. Heldestad V, Linder J, Sellersjo 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):1878–1885.
19. Jones DH, Kilgour RD, Comtois AS. Test-retest reliability of pressure pain threshold measurements of the upper limb and torso in young healthy women. *J Pain*. 2007;8(8):650–656.
20. Knutti IA, Suter MR, Opsommer E. Test-retest reliability of thermal quantitative sensory testing on two sites within the L5 dermatome of the lumbar spine and lower extremity. *Neurosci Lett*. 2014;579:157–162.
21. Lowenstein L, Jesse K, Kenton K. Comparison of perception threshold testing and thermal-vibratory testing. *Muscle Nerve*. 2008;37(4):514–517.
22. 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):547–552.
23. Moloney NA, Hall TM, Doody CM. Reliability of thermal quantitative sensory testing: a systematic review. *J Rehabil Res Dev*. 2012;49(2):191–208.
24. Peltier A, Smith AG, Russell JW, et al. Reliability of quantitative sudomotor axon reflex testing and quantitative sensory testing in neuropathy of impaired glucose regulation. *Muscle Nerve*. 2009;39(4):529–535.

25. 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):655–658.
26. 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.
27. Hirschfeld G, Zernikow B, Kraemer N, et al. Development of somatosensory perception in children: a longitudinal QST-study. *Neuropediatrics*. 2012;43(1):10–16.
28. Pigg M, Baad-Hansen L, Svensson P, Drangsholt M, List T. Reliability of intraoral quantitative sensory testing (QST). *Pain*. 2010;148(2):220–226.
29. 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):548–556.
30. 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):1217–1223.
31. Bland JM, Altman DG. Agreed statistics: measurement method comparison. *Anesthesiology*. 2012;116(1):182–185.
32. Baumgärtner U, Magerl W, Klein T, Hopf HC, Treede RD. Neurogenic hyperalgesia versus painful hypoalgesia: two distinct mechanisms of neuropathic pain. *Pain*. 2002;96(1–2):141–151.
33. Magerl W, Wilk SH, Treede RD. Secondary hyperalgesia and perceptual wind-up following intradermal injection of capsaicin in humans. *Pain*. 1998;74(2–3):257–268.
34. Werner MU, Petersen MA, Bischoff JM. Test-retest studies in quantitative sensory testing: a critical review. *Acta Anaesthesiol Scand*. 2013;57(8):957–963.
35. Mukaka MM. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*. 2012;24(3):69–71.
36. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.
37. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–310.
38. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*. 2005;19(1):231–240.
39. Franz M, Spohn D, Ritter A, Rolke R, Miltner WH, Weiss T. Laser heat stimulation of tiny skin areas adds valuable information to quantitative sensory testing in postherpetic neuralgia. *Pain*. 2012;153(8):1687–1694.
40. Puta C, Schulz B, Schoeler S, et al. Enhanced sensitivity to punctate painful stimuli in female patients with chronic low back pain. *BMC Neurol*. 2012;12:98.
41. Wasner GL, Brock JA. Determinants of thermal pain thresholds in normal subjects. *Clin Neurophysiol*. 2008;119(10):2389–2395.
42. Walton DM, Levesque L, Payne M, Schick J. Clinical pressure pain threshold testing in neck pain: comparing protocols, responsiveness, and association with psychological variables. *Phys Ther*. 2014;94(6):827–837.
43. 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.
44. Geber C, Scherens A, Pfau D, et al. Zertifizierungsrichtlinien für QST-Labore. [Procedure for certification of QST laboratories]. *Schmerz*. 2009;23(1):65–69. German.
45. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol*. 2008;31(4):466–475.
46. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol*. 2006;59(10):1033–1039.
47. Turner JA, Turk DC. The significance of clinical significance. *Pain*. 2008;137(3):467–468.
48. Kottner J, Audige L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Int J Nurs Stud*. 2011;48(6):661–671.

Publikation III

Climate-related differences in temperature pain sensation? Comparison between a mid-European and a sub-Arctic sample

Helen Nothnagel[1], Thomas Weiss[2], Christian Puta[1,3], Gunnvald Kvarstein[4,5], Frauke Musial[6]

[1] Department of Sports Medicine and Health Promotion, Friedrich Schiller University, Jena, Germany
[2] Department of Biological and Clinical Psychology, Friedrich Schiller University, Jena, Germany
[3] Center for Interdisciplinary Prevention of Diseases related to Professional Activities, Friedrich Schiller University, Jena, Germany
[4] Department of Clinical Medicine, UIT, The Arctic University of Norway, Tromsø, Norway
[5] The Multidisciplinary Pain Clinic, Division of Surgical Medicine and Intensive Care, University Hospital of North Norway, Tromsø, Norway
[6] The National Research Center for Complementary and Alternative Medicine (NAFKAM), Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway

Corresponding Author:
Helen Nothnagel[1]
Wöllnitzer Straße 42, Jena, 07749, Germany
E-mail address: helen.nothnagel@uni-jena.de

Abstract

Background. A mid-European (Germany) and a Norwegian (geographically arctic, climatically sub-Arctic) research group cooperate on chronic pain research using the standardized Quantitative Sensory Testing (QST) protocol from the German Research Network on Neuropathic Pain (DFNS). As part of a quality assurance procedure, QST measurements on the dominant hand (dorsum) were compared between the two laboratories.

Methods. The data from both laboratories were collected as part of studies recruiting persons with chronic neck or back pain. Both, participants recruited as healthy controls as well as patients were included into the quality assurance, if i) they scored > 40 for pain intensity on a 0-100 numerical rating scale at the time of the baseline QST measurement, and ii) their QST measurements were normal in reference to the normative data of the DFNS.

Results. The mid-European sample consisted of 18 participants (42.8 ± 11.9 years) and the sub-Arctic sample of 20 participants (41.1 ± 13.2 years). There was no significant group difference for the mechanical tests as well as for the thermal detection thresholds. However, the sub-Arctic sample exhibited lower cold and heat pain thresholds. Possible confounders such as gender, different examiners, or whether the study participants were recruited as healthy controls or patients were excluded.

Conclusion. The sub-Arctic sample was more sensitive to cold and heat pain, even though all participants were within the range of the QST norm data. These findings suggest that potential regional and climate-related differences of a study sample need to be taken into account when determining thermal QST measures.

Introduction

Over the last two decades, Quantitative Sensory Testing (QST), a psychophysical instrument for the investigation of sensory perception, has increasingly been used in pain research and clinical practice (Krassioukov et al., 1999; Chong & Cros, 2004; Cruccu et al., 2010; Pavlakovic & Petzke, 2010; Backonja et al., 2013). In clinical research, the highly standardized QST protocol, established by the German Research Network on neuropathic Pain (DFNS), is one of the most frequently used tools to determine detection and pain thresholds for thermal and mechanical stimuli (Rolke et al., 2006a; Rolke et al., 2006b; Magerl et al., 2010; Maier et al., 2010). One of the major advantages of using this particular QST protocol, is the availability of a reference database with age- and gender-matched normative data from healthy volunteers, with norm values for the face, hand, and foot (Magerl et al., 2010). Additional reference values for the upper and lower back, stratified according to gender, age and location of stimuli, were published in 2014 (Pfau et al., 2014). The reliability and reproducibility of the standardized QST protocol reveal a good test-retest and inter-observer reliability within 2 days (Geber et al., 2011). Moreover, a more recent study found that most parameters of the QST were reliable and reproducible over 10 weeks (Nothnagel et al., 2017).

The elaborated protocol with its high standardization, the likewise sophisticated training program of the DFNS for persons conducting the QST protocol, and the fact, that age- and gender-related

norm data are available, further makes this protocol a suitable instrument for multicenter studies and cooperations on pain research between different laboratories. Therefore, this protocol was chosen for a scientific collaboration on pain research between the University of Jena (Germany) and at UiT, The Arctic University of Norway (Tromsø, Norway).

As the two research groups aim to cooperate on future multicenter clinical studies, there were two rationales for a post-hoc analysis: a) Even though a highly standardized QST protocol was used by specially trained investigators, we aimed to ensure that there was no systematic difference between the laboratories in the way the protocol was implemented and conducted. b) Temperature-related measurements are part of most standard QST protocols, and this is also true for the DFNS QST protocol. The inhabitants of Tromsø, with a geographic location 350 km north of the polar circle, are exposed to long winters with rather cold temperatures even in the urban environments. To our knowledge, the possibility of systematic influences of climate on the DFNS QST protocol have not yet been investigated. Thus, we wanted to exclude potential climate related systematic influences on the outcomes of the protocol, in order to avoid risk of bias in future multicenter studies using this particular QST protocol.

Materials & Methods

Available data material

The first two studies on QST and chronic pain within this framework were performed between 2014 and 2019. The first study tested the reproducibility of QST measures and was conducted at the Department of Sports Medicine and Health Promotion of the University of Jena (Jena, Germany). For this study, healthy volunteers and chronic low back pain patients were included (the data from healthy volunteers are presented in Nothnagel et al. (2017)). The second study was conducted at the National Research Center for Complementary and Alternative Medicine (NAFKAM), Department of Community Medicine, the Arctic University of Norway (Tromsø, Norway) and aimed to investigate non-pharmacological treatment of patients with chronic, non-specific neck pain. It turned out that many of the participants had fallen out of the main inclusion criterion (baseline pain intensity > 40 on a 100-point numerical rating scale (NRS, 0 = “no pain”, 100 = “worst pain imaginable”)) by the time of first QST measurement. This phenomenon was related to challenges with recruitment and inclusion procedures already discussed in Nothnagel et al. (2019).

Inclusion criteria

Data were included for the post-hoc analysis, if i) the participants rated mean pain intensity at the time of QST measurement to be ≤ 40 on a 0-100 numerical rating scale, and ii) the QST data were within the healthy range of the normative data of the DFNS. Data of study participants, recruited as healthy controls and as pain patients but exhibiting normal data at the time of measurement were thus included in this analysis.

112 *Participants*

113 QST measurements from a total of 38 study participants, 18 from Jena, Germany (10 males and 8
114 females), and 20 from Tromsø, Norway (1 male and 19 females; recruited as pain patients
115 N=19), fulfilled the inclusion criteria.

116

117 *Ethics*

118 The present post-hoc analysis compared completely anonymized data drawn from two different
119 studies. The first study was conducted at the department of Sports Medicine and Health
120 Promotion at Friedrich Schiller University Jena, Germany. The local ethics committee of the
121 University of Jena approved the study protocol and the study was performed in accordance with
122 the Helsinki Declaration. All volunteers provided written informed consent prior to participation.
123 This work was supported by the federal Ministry of Education and Research BMBF (01EC1003,
124 01EC1010).
125
126 The second study was conducted at the National Research Center for Complementary and
127 Alternative Medicine (NAFKAM), Department of Community Medicine, the Arctic University
128 of Norway (Tromsø, Norway). This study was approved by the regional ethics committee
129 (Regional committees for medical and health research ethics, REK Nord registration number
130 2014/1105). Participants were informed about the purpose of the study, risk/benefit profile of the
131 interventions, and the study itself. All participants were free to withdraw from participation at
132 any point if they wished to. All participants provided written informed consent prior to
133 participation. The study was performed in accordance with the declaration of Helsinki.
134
135 All identifiable personal data were destroyed after completing the two studies. Since only
136 completely anonymized data from both studies were combined in this post-hoc analysis, no
137 ethical approval for this post-hoc analysis was needed.
138
139

140 *Experimental procedure*

141 The QST examination was conducted in accordance with the recommendations and protocol of
142 the DFNS (Rolke et al., 2006a; Magerl et al., 2010) in a quiet, closed room and with test
143 instructions provided by the DFNS. Room and skin temperatures were recorded (see Table 1).
144 The data from Tromsø were collected by two different examiners (the first author HN and BD,
145 see Acknowledgement), who both were trained by DFNS. Only the data collected on the dorsum
146 of the dominant hand were included in the post-hoc analysis.

147 In the laboratory in Jena, Germany the thermal stimulator Thermal Sensory Analyzer II (TSA
148 2001-II, Medoc, Israel; stimulation area of 30*30 mm), and in the laboratory in Tromsø, Norway
149 the Modular Sensory Analyzer (MSA, Somedic, Sweden; stimulation area of 25*50 mm) was
150 used. Both thermal stimulators are approved and recommended for the conduction of the DFNS
151 QST protocol.

We determined the following thermal detection and pain thresholds: Cold detection threshold (CDT), warm detection threshold (WDT), thermal sensory limen procedure (TSL) of alternating cold and warm stimuli, cold pain threshold (CPT), and heat pain threshold (HPT) (Rolke et al., 2006a; Rolke et al., 2006b). All thresholds were obtained with ramped stimuli of 1 °C/s which ceased when the participant pressed a stop button. Baseline temperature was 32 °C with a lower cut-off temperature at 0 °C (3.9 °C by MSA) and upper cut-off temperature at 50 °C. The mean threshold temperature of three consecutive measurements was calculated.

Mechanical detection threshold (MDT) for touch was assessed using a standardized set of modified von Frey filaments (diameter 0.5 mm, Optihair$_2$-Set Marstock Nervtest, Schriesheim, Germany) which exerts forces between 0.25 and 512 mN (factor 2 progression). Using the "method of limits", the threshold presented is the geometric mean of five consecutive series with ascending and descending stimulus intensities (Baumgärtner et al., 2002; Rolke et al., 2006a).

Mechanical pain threshold (MPT) was measured using a set of standard pinprick stimulators (cylindrical tip, 250 µm tip diameter) with fixed stimulus intensities exerting forces of 8, 16, 32, 64, 128, 256, and 512 mN (MRC Systems GmbH, Heidelberg, Germany). The stimulation was applied in ascending order until the first perception of sharpness was detected. MPT was determined using the "method of limits". The threshold value presented is the geometric mean of five consecutive series with ascending and descending stimuli intensities.

Pain induced by punctate mechanical stimuli was measured using the same standard pinprick stimulators as for MPT. To obtain mechanical pain sensitivity (MPS) towards pinprick-evoked pain, all seven pinprick stimuli were applied in balanced order, five times for each stimulus. MPS was defined as the geometric mean of the given stimuli. To avoid effects of sensitization or fatigue, the successive stimuli were not applied to the same spot of skin, but a few millimeters away from the previously stimulated spot. Following each stimulus, participants were asked to rate pain intensity for each stimulus on a NRS with 0 indicating "no pain at all" and 100 indicating "worst pain imaginable". Pain to light touch was assessed by lightly stroking the participant with a cotton wisp (~3 mN), a Q-tip affixed to an elastic strip (~100 mN), and a soft make-up brush (~200-400 mN). The set of the three light tactile stimulators were intermingled with the pinprick stimuli in balanced order. If the stroking stimuli were perceived as painful, participants were asked to use the same NRS (0-100) to provide a rating for the extent of the pain perceived.

The perceptual correlate of temporal pain summation to repetitive pinprick stimuli (Wind-up ratio, WUR) was assessed by a series of ten pinprick stimuli (256 mN) with 1 Hz repetition rates. Again, the participants were asked to give a pain rating to define the pain at the end of the stimuli series using the 0-100 NRS. The pain ratings to single pinprick stimulation were compared to those of ten repeated stimuli. To determine the WUR, the ratio of the mean pain rating of the series divided by the mean pain rating to a single stimulus was calculated after five trials.

Vibration detection threshold (VDT) was measured using a standardized Rydel-Seiffer graduated tuning fork (64 Hz, 8/8 scale) that was placed over a bony protrusion (hand: ulnar styloid process) (Rolke et al., 2006a; Rolke et al., 2006b; Magerl et al., 2010). Participants reported the disappearance of the vibration. VDT was determined as the average of three consecutive measurements of the amount of time to disappearance.

Pressure pain threshold (PPT) was assessed over a muscle on the test areas (hand: pollicis muscles/thenar eminence) using a pressure gauge device (FDN200, Wagner Instruments, Greenwich, CT, USA) with a probe area of 1 cm^2 that exerts pressure up to 2000 kPa. The PPT was determined by three series of ascending stimulus intensities, each applied as an increasing ramp of 50 kPa/s.

Statistical analysis

The mean of consecutive QST measurements for each single subject were calculated using Excel (Office, 2013, Microsoft, USA). These results were summarized in a single QST report form sheet for each subject and used for further statistical analysis. In accordance with standard QST procedure, all QST variables, with the exception of CPT, HPT, and VDT, were logarithmically transformed (base 10) to achieve a (secondary) normal distribution (Rolke et al., 2006a). To ensure that no values were lost due to zero rating (Magerl, Wilk & Treede, 1998), a small constant (0.1) was added to pain ratings for pinprick (MPS) prior to log-transformation. Normal distribution was tested by Shapiro-Wilk test.

Due to the use of two different thermal stimulators (TSA and MSA) with divergent lower cut-off temperatures (0 °C with TSA and 3.9 °C with MSA), the CPT data were modified before analyzing. All CPT values below 3.9 °C from the German study sample were modified to exactly 3.9 °C. This strategy is statistically conservative, since it reduces potential differences between the two samples and thus avoids overestimation of potential group differences.
As recommended by the protocol of the DFNS (Rolke et al., 2006a; Rolke et al., 2006b), the QST data of each participant were transformed into a standard normal distribution (z-transformation: zero mean, unit variance) by the following expression:

$$z\text{-score} = (\text{value}_{\text{participant}} - \text{mean}_{\text{DFNS reference}})/\text{SD}_{\text{DFNS reference}}$$

Analysis of variance (ANOVA) was conducted for group differences. We used the between-subject factor sample (mid-European sample and sub-Arctic sample) and the within-subject factors QST modalities (CDT, WDT, TSL, CPT, HPT, PPT, MPT, MPS, WUR, MDT, and VDT). Results were corrected for violations of sphericity using the Greenhouse-Geisser ε-correction for degrees of freedom. Post-hoc analyses was performed using independent sample t-tests. Significance level was adjusted according to Bonferroni with a resulting p-value of $p \leq 0.0045$. Since a direct comparison of QST subtests will result in differences based on the nature of the subtest, only the between-subject factor sample and interactions of this factor with the factor QST modalities were interpreted.

Analysis of covariance (ANCOVA) was used for the further assessment of the NRS as a confounding factor. Frequencies were compared by means of Chi^2-test. With regard to the low number of data points correlations were calculated non-parametrically with Kendall-Tau-b. All statistical calculations were performed using SPSS Statistics 22 (IBM Deutschland GmbH, Germany).

Results

Descriptive characteristics

Table 1 shows the descriptive characteristics of all participants (N=38). From Jena, Germany (mid-European sample), 18 participants (10 males and 8 females, mean age 42.8 ± 11.9 years) and from Tromsø, Norway (sub-Arctic sample), 20 participants (1 male and 19 females, mean age 41.1 ± 13.2 years) were evaluated. Of these, 34 candidates (89 %) were right-handed, while 4 (11 %) were left-handed. Besides an asymmetric gender distribution, no significant differences were found between the groups with regard to age or anthropometric data (weight, height, and body mass index). The NRS ranged from 0 to 40 on a 100-point numerical rating scale in both groups and did not differ between the groups, nor did skin and room temperatures.

Table 1 should be inserted here

QST data

Table 2 shows the thermal and mechanical QST data of both groups (mid-European and sub-Arctic sample). Note that all QST values represent normalized z-scores.

Table 2 should be inserted here

ANOVA revealed a significant interaction between QST modalities and sample ($F(7.066, 254.393)=3.451$, $p=0.001$). Even though the p-value was $p=0.055$, the main effect for the factor sample ($F(1,36)=3.938$) did not reach statistical significance. Post-hoc analysis revealed no differences between the two groups (Jena, Germany and Tromsø, Norway) with regard to their thermal detection thresholds (CDT, $p=0.200$; WDT, $p=0.088$; TSL, $p=0.509$). The analysis of the thermal pain thresholds (CPT and HPT), however, revealed significant differences between the two groups. The Tromsø sample exhibited a lower cold pain threshold (CPT) than the study sample in Jena, Germany ($p=0.001297$). A similar picture emerged with regard to the heat pain threshold (HPT). The sub-Arctic sample (Tromsø, Norway) exhibited a lower HPT than the mid-European sample (Jena, Germany) ($p=0.000189$). With regard to all mechanical thresholds (PPT, MPT, MPS, WUR, MDT, and VDT), no significant differences were seen between the mid-European and the sub-Arctic sample.

Figure 1 shows the z-transformed QST data of Table 1 for both groups (mid-European and sub-Arctic sample).

Figure 1 should be inserted here

Possible confounding factors and sources of bias

Gender distribution

The sample of this post-hoc analysis revealed a skewed gender distribution, with only 11 males and 27 females included (p=0.000601). This was in particular the case for the Tromsø population, which included only one male participant. Gender had no influence on the mechanical thresholds.
To determine whether the differences in thermal pain thresholds (CPT and HPT) between the German and Norwegian samples were due to different gender distribution between the groups, we repeated the statistical analysis with female participants only (N=27). Table 3 shows the thermal and mechanical detection and pain thresholds as z-scores in both groups.
ANOVA revealed a significant interaction between QST modalities and sample ($F(6.624, 165.590)=2.652$, $p=0.014$). There was no significant main effect of the factor sample ($F(1,25)=2.934$, $p=0.099$). Post-hoc analysis revealed that the thermal pain thresholds were significantly lower with low p-values for CPT ($p=0.015$) and HPT ($p=0.023$). Thus, this result is similar to the total sample, including men and women.

Table 3 should be inserted here

Study participants, recruited as pain patients
One possible reason for differences in thermal pain thresholds could be the fact, that the sample in Tromsø included more study participants recruited as "pain patients". Even though all participants were technically pain free with an NRS score ≤ 40 and sensory skin function "normal" in the understanding, that their QST values scored within the normal range of the age and gender matched norms available for the DFNS QST protocol. The self-identification as "pain patient" and conditioning from a history of pain may still constitute a bias. This may in particular be true, because the study participants in Tromsø were recruited into a neck pain study and it cannot be ruled out that neck pain may influence pain related measurements of the hand. Thus, other possible differences that might influence on thermal pain perception (CPT and HPT) was further investigated. Low pain intensity (NRS ≤ 40) was the main inclusion criterion for inclusion in the study in Jena and in Tromsø, and the following analysis therefore focused on potential differences in the NRS, between the participants, recruited as pain patients or healthy controls and its possible correlation to CPT and HPT.

Figure 2 shows NRS scores of the mid-European and the sub-Arctic sample, divided according to whether the study participants were recruited as pain patients, or healthy volunteers. Table 4 shows the numbers within the groups and the NRS values. While the sub-Arctic sample included only one person recruited as healthy control, the mid-European sample contained 10 persons recruited as healthy control, but the NRS scores for the two samples did not differ. However, if only the persons recruited as patients were included in the comparison between the two samples (as healthy controls scored "0" on the NRS), the NRS values of the sub-Arctic sample were actually lower (p=0.001) (Table 4).

Figure 2 should be inserted here

Table 4 should be inserted here

Within each geographical group the correlations between NRS scores and z-transformed CPT and HPT values (See Table 5 and Fig. 3) were weak for participants recruited as pain patients (Fig. 3), and only significantly positive between NRS and HPT for the total sample. Therefore, even though there was a mean difference in mean pain intensity between the geographical samples when only persons recruited as patients were included, there was little systematic covariation between NRS and CPT or HPT. Figure 3 depicts the scatterplots of the correlations between the NRS scores and the CPT and HPT respectively for those participants recruited as pain patients.

Table 5 should be inserted here

Figure 3 should be inserted here

Nonetheless, in order to estimate the effect of a potential covariation of the NRS with CPT and HPT values, an ANCOVA with the NRS as a covariate was calculated (See Table 6). The results show, that even when the covariation of the NRS with the CPT and HPT values is extracted, the thermal pain threshold differences between the sub-Arctic and the mid-European sample remain.

Table 6 should be inserted here

Examiners' comparison

The data were collected by two examiners (HN and BD, see acknowledgement) and both were trained by DFNS. HN measured all participants from Jena, Germany (100 %, N=18) and 65 % of participants from Tromsø, Norway (N=13). BD measured 7 participants from Tromsø, Norway (35 %). The QST values of sub-Arctic sample (N=20) of both examiners (HN and BD) were compared by the independent sample t-tests. There were no relevant differences (Table 7).

Table 7 should be inserted here

Discussion

The aim of the analysis presented above was twofold: a) To establish a base for future multicenter studies, we aspired to implement a highly standardized QST protocol in such a way that there was no concern about possible sources of bias between the laboratories, b) Since comprehensive QST protocols include temperature thresholds, we wanted to investigate potential differences regarding the temperature related QST parameters of the DFNS protocol.

No differences between the laboratories were found for the mechanical tests (PPT, MPT, MPS, WUR, MDT, and VDT). The same held true for the thermal detection thresholds (CDT, WDT, and TSL). In contrast to these findings, the thresholds for cold pain (CPT) and heat pain (HPT) were different between the two study samples: The subarctic sub-Arctic sample exhibited greater pain sensitivity to cold as well as to heat (Table 1; Fig. 1) compared with the mid-European sample, living in a significantly different climate zone. We have not identified relevant publications investigating such a phenomenon and it has to be taken into account that a post-hoc analysis is more vulnerable to bias than a systematically planned prospective study although we were able to control for relevant sources of bias.

Furthermore, it should be born in mind that QST measures are psychophysical measures based on subjective reports. A sub-Arctic population may possess an adaptive and learned behavior to low temperatures which pose a considerable health risk several months of the year. Such a learned behavior would therefore be highly adaptive. If this is correct, investigators using QST need to be aware of potential climate-related biases in QST.

Some information about regional or ethnic differences in thermal perception is available in the literature. Earlier data on pain detection thresholds of women from various ethnic groups (Yankee, Irish, Jewish, and Italian housewives), using electric shock to trigger a pain stimulus, showed no differences between the groups (Sternbach & Tursky, 1965). However, more recent studies have revealed cultural and ethnic influences on pain perception and pain thresholds (Zatzick & Dimsdale, 1990; Edwards & Fillingim, 1999; Edwards, Fillingim & Keefe, 2001; Dawson & List, 2009). In a study consisting of young healthy adults, no differences were found in thermal pain thresholds or ratings of thermal pain intensity. Lower thermal pain tolerances, however, were found in African-Americans, who rated thermal heat stimuli as more unpleasant (Edwards & Fillingim, 1999). Further studies, investigating differences in pain experience in African-Americans and Caucasians, reported lower heat pain thresholds and tolerance in African-Americans, as well as significantly lower pressure pain tolerances (Edwards, Fillingim & Keefe, 2001). Sheffield also observed significantly higher ratings of thermal pain unpleasantness and marginally higher ratings of pain intensity in African-Americans (Sheffield et al., 2000). Gender and professional status of the experimenter seem to exert an influence on thermal pain reports (Levine & de Simone, 1991; Kállai, Barke & Voss, 2004; Aslaksen et al., 2007). All in all, a variety of biological, social, and psychological mechanisms are proposed to be responsible for observed differences in pain perception and pain tolerance, while geographic and climatic influences have received little attention. We find the cultural and ethnic differences discussed above, not liable to play a role in the comparison between our mid-European and a sub-Arctic study samples, as both can be characterized as Caucasian.

Limitations

Possible sources of bias in our study include the fact, that i) the participants were partly pain patients, although they did not fulfill the pain intensity criterion at the time of QST measurement, ii) gender was not equally distributed between the two samples in the two laboratories, iii) two different investigators conducted the QST in the Norwegian sample.
A detailed analysis of possible differences between the participants, recruited as pain patients, and those, recruited as healthy controls revealed an unequal distribution between the two study locations, but the total samples were comparable regarding NRS values and a correlation analysis revealed no interdependence of the NRS scores with neither CPT nor HPT. When only persons, recruited as pain patients, were included in the comparison the sub-Artic subsample revealed lower NRS values. We find it unlikely that these lower NRS scores explained the lower CPT and HPT scores in the sub-Arctic sample. We would expect that a sample which is more sensitive to cold and heat pain report higher general pain values, not lower.
Secondary analyses revealed, that the results still hold, also when only women were included and a difference between the two investigators in Tromsø were not demonstrated for any of the measurements.

One possible mediator of the differences in temperature related pain thresholds could be the outside temperature during the study period as there were seasonal differences between the time frames of data collection of the two studies: The study in Jena was conducted from March to October, including the warm summer season, while the study in Tromsø was conducted between October and April, which includes the polar night and the coldest period of the year. It is difficult to estimate how much the outside temperatures might influence on the results. However, as Table 1 shows, there were no differences in skin or room temperature between the two laboratories. In addition, such an influence would just emphasize, how important it is to systematically investigate the possible influence of environmental conditions, such as different climate zones. To our knowledge, no data about possible seasonal as well as climate related differences in QST profiles are available.

There is one additional caveat to the interpretation of the results. A recent report on chronic pain released by the Norwegian Institute of Public Health (NIPH) states, that Norwegian citizens, compared to other countries, suffer from an unusual high degree of chronic pain, which affects about 30 per cent of the adult Norwegian population. It is the most common reason for long-term sick leave and disability benefit (Breivik et al., 2006; Nielsen et al., 2019). These results are difficult to interpret, as the corresponding numbers in other Nordic countries are considerably lower. How these clinical data are to be interpreted and how they relate to the QST results presented here and the climate zone remains to be elucidated.

A limitation with regard to the interpretation of these data is, that it remains unclear, whether these differences in thermal threshold are due to a physiologically based shift of these thresholds, or whether the differences are due to different subjective criteria, when to react. The latter has a certain likelihood, which could represent learned behavior which helps inhabitants of a subarctic climate zone to adjust to the environment. Applying a signal detection threshold paradigm (SDT) would give insight into this question and should be considered when investigating this phenomenon further (Abdi, 2007).

Conclusion

The comparison of QST measurements between a sub-Arctic and a mid-European study sample, conducted according to the highly standardized protocol of the DFNS, revealed robust and substantial differences for the thermal pain thresholds (CPT and HPT), while all other QST parameters were comparable. The possible mechanism behind this effect remains unclear and there are several potential explanations for the effect.

With regard to multicenter studies on pain, these data indicate, that environmental conditions, such as climate, may impose a risk of bias even when using such a highly standardized protocol as the QST protocol of the DFNS. Further systematic investigation of the effect in prospective studies is urgently needed, in order to further elucidate climate related differences in temperature perception and to be able to estimate the size of these effects in clinical trials on pain.

Acknowledgments

We would like to thank Brit Drageset, NAFKAM, The Arctic University of Norway, UiT, Tromsø, for her technical support during the QST measurements.

Author Contributions

GK, TW, CP, and FM conceived and planned the designs of the studies. F.M. supervised the implementation of the study. GK contributed to the interpretation of the results. HN carried out the experiments and wrote the manuscript with support from FM. All authors discussed the results and contributed to the final manuscript.

Disclosure

All authors declare that there is no conflict of interest.

References

Abdi H. 2007. Signal detection theory (SDT). *Encyclopedia of measurement and statistics*: 886-889.

Aslaksen PM, Myrbakk IN, Høifødt RS, Flaten MA. 2007. The effect of experimenter gender on autonomic and subjective responses to pain stimuli. *Pain* 129(3): 260-268.

Backonja MM, Attal N, Baron R, Bouhassira D, Drangholt M, Dyck PJ, Edwards RR, Freeman R, Gracely R, Haanpaa MH, Hansson P, Hatem SM, Krumova EK, Jensen TS, Maier C, Mick G, Rice AS, Rolke R, Treede RD, Serra J, Toelle T, Tugnoli V, Walk D, Walalce MS, Ware M, Yarnitsky D, Ziegler D. 2013. Value of quantitative sensory testing in neurological and pain disorders: NeuPSIG consensus. *Pain* 154(9): 1807-1819.

Baumgärtner U, Magerl W, Klein T, Hopf HC, Treede RD. 2002. Neurogenic hyperalgesia versus painful hypoalgesia: two distinct mechanisms of neuropathic pain. *Pain* 96(1-2); 141-151.

Breivik H, Collett B, Ventafridda V, Cohen R, Gallacher D. 2006. Survey of chronic pain in Europe: Prevalence, impact on daily life, and treatment. *Eur J Pain* 10(4): 287-287.

Chong PS, Cros DP. 2004. Technology literature review: quantitative sensory testing. *Muscle Nerve* 29(5): 734-747.

Cruccu G, Sommer C, Anand P, Attal N, Baron R, Garcia-Larrea L, Haanpaa M, Jensen TS, Serra J, Treede RD. 2010. EFNS guidelines on neuropathic pain assessment: revised 2009. *Eur J Neurol* 17(8): 1010-1018.

Dawson A, List T. 2009. Comparison of pain thresholds and pain tolerance levels between Middle Easterners and Swedes and between genders. *Journal of oral rehabilitation* 36(4): 271-278.

Edwards RR, Fillingim RB. 1999. Ethnic differences in thermal pain responses. *Psychosomatic medicine* 61(3): 346-354.

Edwards CL, Fillingim RB, Keefe F. 2001. Race, ethnicity and pain. *Pain* 94(2): 133-137.

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 RD. 2011. 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* 152(3): 548-556.

Kállai I, Barke A, Voss U. 2004. The effects of experimenter characteristics on pain reports in women and men. *Pain* 112(1-2): 142-147.

Krassioukov A, Wolfe DL, Hsieh JT, Hayes KC, Durhamn CE. 1999. Quantitative sensory testing in patients with incomplete spinal cord injury. *Arch Phys Med Rehabil* 80(10): 1258-1263.

Levine FM, de Simone LL. 1991. The effects of experimenter gender of pain report in male and female subjects. *Pain* 44(1): 69-72.

Magerl W, Wilk SH, Treede RD. 1998. Secondary hyperalgesia and perceptual wind-up following intradermal injection of capsaicin in humans. *Pain* 74(2-3); 257-268.

Magerl W, Krumova EK, Baron R, Tölle T, Treede RD, Maier C. 2010. Reference data for quantitative sensory testing (QST): refined stratification for age and a novel method for statistical comparison of group data. *Pain* 151(3): 598-605.

Maier C, Baron R, Tölle TR, Binder A, Birbaumer N, Birklein F, Gierthmühlen J, Flor H, Geber C, Huge V, Krumova EK, Landwehrmeyer GB, Magerl W, Maihöfner C, Richter H, Rolke R, Scherens A, Schwarz A, Sommer C, Tronnier V, Üçeyler N, Valet M, Wasner G, Treed RD. 2010. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): somatosensory abnormalities in 1236 patients with different neuropathic pain syndromes. *Pain* 150(3): 439-450.

Nielsen C, Steingrimsdottir Ö, Handal M, Skurtveit S. 2019. Chronic pain. Norwegian Institute of Public Health.

Nothnagel H, Puta C, Lehmann T, Baumbach P, Menard MB, Gabriel B, Gabriel HHW, Weiss T, Musial F. 2017. How stable are quantitative sensory testing (QST) measurements over time? Report on 10-week reliability and agreement of results in healthy volunteers. *Journal of Pain Research* 10: 2067-2078.

Nothnagel H, Menard MB, Kvarstein G, Norheim AJ, Weiss T, Puta C, Mist SD, Musial F. 2019. Recruitment and inclusion procedures as "pain killers" in clinical trials? *Journal of Pain Research* 12: 2027-2037.

Pavlakovic G, Petzke F. 2010. The role of Quantitative sensory testing in the evaluation of musculoskeletal pain conditions. *Curr Rheumatol Rep* 12: 455-461.

543 Pfau DB, Krumova EK, Treede RD, Baron R, Toelle T, Birklein F, Eich W, Geber C, Gerhardt
544 A, Weiss T, Magerl W, Maier C. 2014. Quantitative sensory testing in the German
545 Research Network on Neuropathic Pain (DFNS): reference data for the trunk and
546 application in patients with chronic postherpetic neuralgia. *Pain* 155(5): 1002-1015.

547 Rolke R, Baron R, Maier CA, Tölle TR, Treede RD, Beyer A, Binder A, Birbaumer N, Birkleich
548 F, Bötefür IC, Braune S, Flor H, Huge V, Klug R, Landwehrmeyer GB, Magerl W,
549 Maihöfner C, Rolko C, Schaub C, Scherens A, Sprenger T, Valet M, Wasserka B. 2006a.
550 Quantitative sensory testing in the German Research Network on Neuropathic Pain
551 (DFNS): standardized protocol and reference values. *Pain* 123(3): 231-243.

552 Rolke R, Magerl W, Campbell KA, Schalber C, Caspari S, Birklein F, Treede RD. 2006b.
553 Quantitative sensory testing: a comprehensive protocol for clinical trials. *Eur J Pain*
554 10(1): 77-88.

555 Sheffield D, Biles PL, Orom H, Maixner W, Sheps DS. 2000. Race and sex differences in
556 cutaneous pain perception. *Psychosomatic Med* 62: 517-523.

557 Sternbach RA, Tursky B. 1965. Ethnic differences among housewives in psychophysical and
558 skin potential responses to electric shock. *Psychophysiology* 1(3): 241-6.

559 Zatzick DF, Dimsdale JE. 1990. Cultural variations in response to painful stimuli. *Psychosomatic*
560 *Med* 52: 544-557.

Figure 1

z-transformed QST variables.

Z-transformed QST variables presented as mean with standard error of the mean (SEM) for the dominant hand of mid-European sample (N=18) and sub-Arctic sample (N=20). **Notes:** Mean and SEM of QST z-scores are plotted for both groups: The triangles represent the data of the mid-European sample and the circles represent the data of the sub-Arctic sample. The horizontal dashed lines mark the 95 % confidence interval of the distribution of healthy subjects (healthy range of normative QST data). *Significant differences between the mid-European sample and sub-Arctic sample, Bonferroni adjusted p≤0.0045. **Abbreviations:** QST = quantitative sensory testing; CDT = cold detection threshold; WDT = warm detection threshold; TSL = thermal sensory limen; CPT = cold pain threshold; HPT = heat pain threshold; PPT = pressure pain threshold; MPT = mechanical pain threshold; MPS = mechanical pain sensitivity; WUR = wind-up ratio; MDT = mechanical detection threshold; VDT = vibration detection threshold; SEM = standard error of the mean.

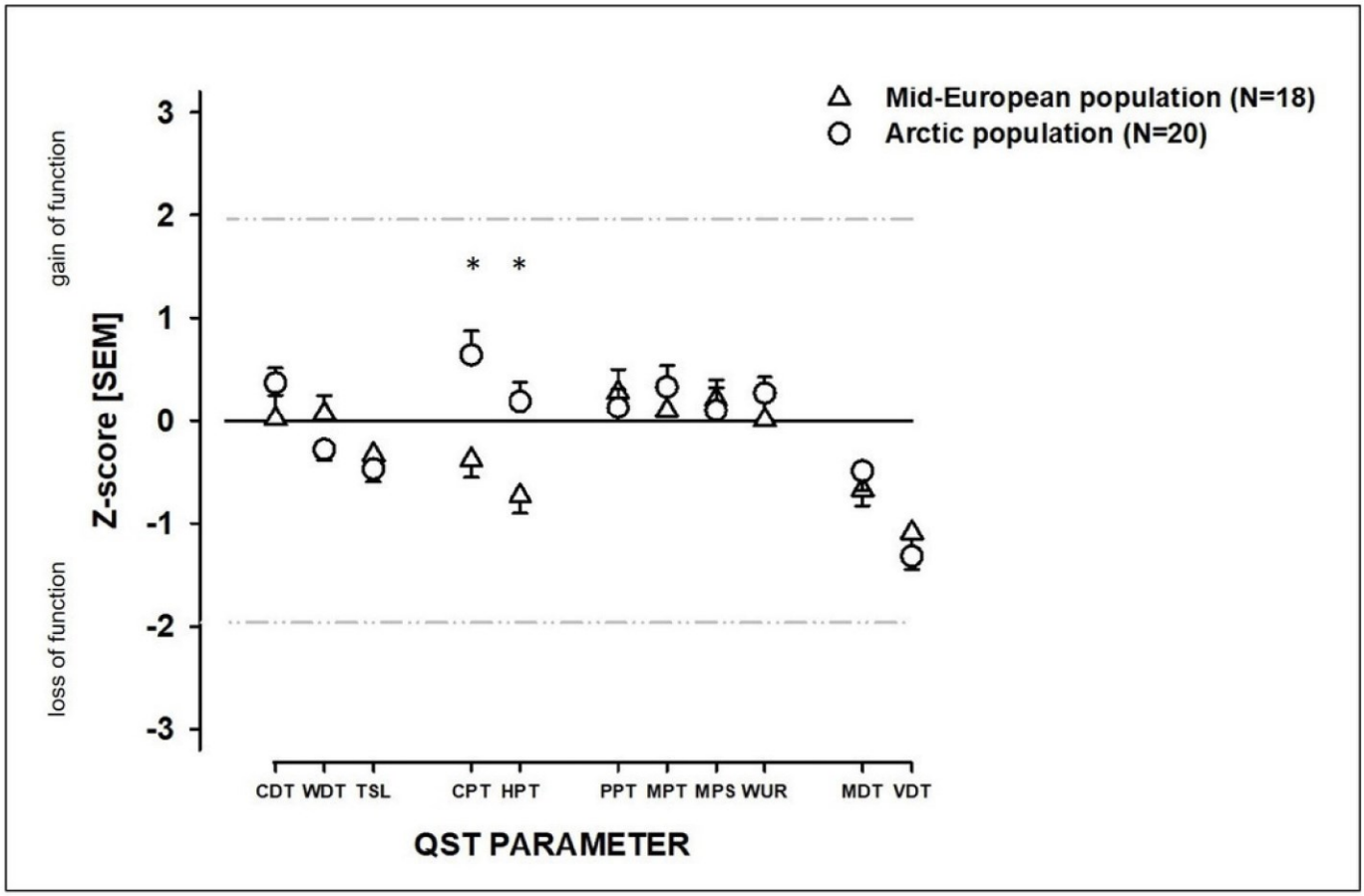


Figure 1: z-transformed QST variables.
Z-transformed QST variables presented as mean with standard error of the mean (SEM) for the dominant hand of mid-European sample (N=18) and sub-Arctic sample (N=20).
Notes: Mean and SEM of QST z-scores are plotted for both groups: The triangles represent the data of the mid-European sample and the circles represent the data of the sub-Arctic sample. The horizontal dashed lines mark the 95 % confidence interval of the distribution of healthy subjects (healthy range of normative QST data). *Significant differences between the mid-European sample and sub-Arctic sample, Bonferroni adjusted p≤0.0045.
Abbreviations: QST = quantitative sensory testing; CDT = cold detection threshold; WDT = warm detection threshold; TSL = thermal sensory limen; CPT = cold pain threshold; HPT = heat pain threshold; PPT = pressure pain threshold; MPT = mechanical pain threshold; MPS = mechanical pain sensitivity; WUR = wind-up ratio; MDT = mechanical detection threshold; VDT = vibration detection threshold; SEM = standard error of the mean.

Figure 2

Pain intensity (NRS) scores of the mid-European (N=18) and sub-Arctic sample (N=20)

Notes: The numeric rating scale (NRS) ranges from 0-100 (0= “no pain”, 100= “worst pain imaginable” are plotted for both groups: The filled triangles represent data of mid-European participants recruited as healthy controls (N=10) while open triangles represent data of mid-European participants recruited as pain patients (N=8). The filled circles represent data of subarctic participants recruited as healthy controls (N=1) while the open circles represent data of sub-Arctic participants recruited as pain patients (N=19). The horizontal dashed line marks the main inclusion criterion of NRS ≤ 40. **Abbreviations:** NRS = numerical rating scale.

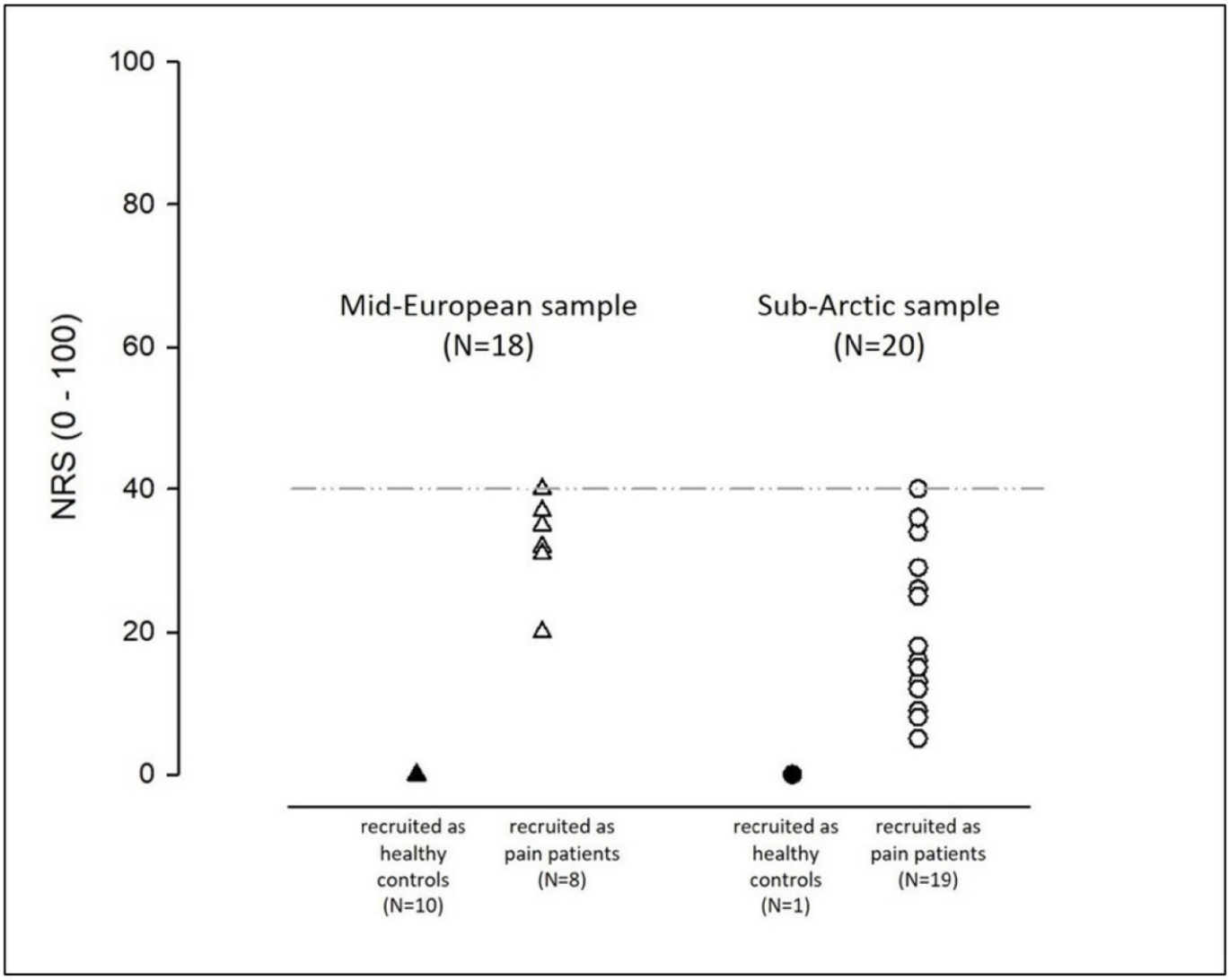


Figure 2: Pain intensity (NRS) scores of the mid-European (N=18) and sub-Arctic sample (N=20).
Notes: The numeric rating scale (NRS) ranges from 0-100 (0= "no pain", 100= "worst pain imaginable" are plotted for both groups: The filled triangles represent data of mid-European participants recruited as healthy controls (N=10) while open triangles represent data of mid-European participants recruited as pain patients (N=8). The filled circles represent data of subarctic participants recruited as healthy controls (N=1) while the open circles represent data of sub-Arctic participants recruited as pain patients (N=19). The horizontal dashed line marks the main inclusion criterion of NRS ≤ 40.
Abbreviations: NRS = numerical rating scale.

Figure 3

Correlations between pain intensity scores and CPT and HPT.

Correlations between pain intensity scores (NRS, 0-100) and A) CPT z-scores values and B) HPT z-scores values of participants recruited as pain patients in the mid-European (N=8) and sub-Arctic sample (N=19). **Notes:** Numeric rating scale ranging from 0-100 (0= “no pain”, 100= “worst pain imaginable”) are plotted for both groups against the A) CPT (z-scores) and B) HPT (z-scores): The triangles represent the data of the participants recruited as pain patients of mid-European sample (N=8). The circles represent the data of participants recruited as pain patients of sub-Arctic sample (N=19). The horizontal dashed lines mark the main inclusion criterion of NRS ≤ 40. **Abbreviations:** NRS = numerical rating scale; CPT = cold pain threshold; HPT = heat pain threshold.

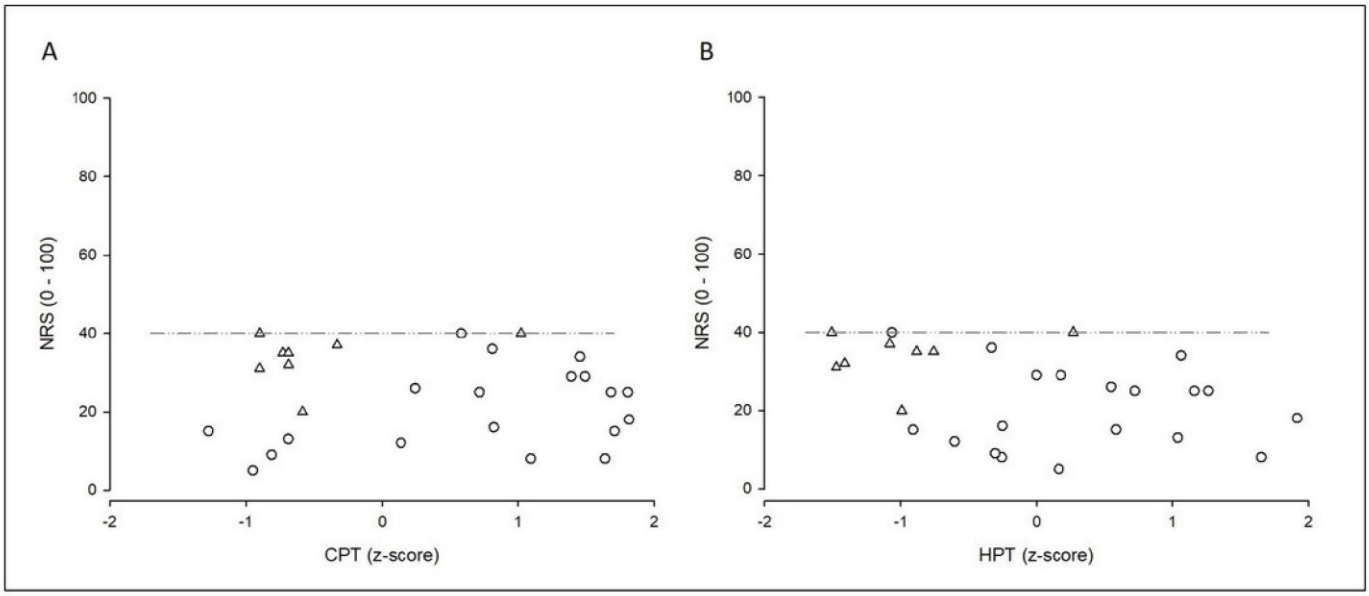


Figure 3: Correlations between pain intensity scores and CPT and HPT.
Correlations between pain intensity scores (NRS, 0-100) and A) CPT z-scores values and B) HPT z-scores values of participants recruited as pain patients in the mid-European (N=8) and sub-Arctic sample (N=19).
Notes: Numeric rating scale ranging from 0-100 (0= “no pain”, 100= “worst pain imaginable”) are plotted for both groups against the A) CPT (z-scores) and B) HPT (z-scores): The triangles represent the data of the participants recruited as pain patients of mid-European sample (N=8). The circles represent the data of participants recruited as pain patients of sub-Arctic sample (N=19). The horizontal dashed lines mark the main inclusion criterion of NRS ≤ 40.
Abbreviations: NRS = numerical rating scale; CPT = cold pain threshold; HPT = heat pain threshold.

1 Table 1. Descriptive characteristics of the total sample (N=38) being analyzed.

		Mid-European sample	**Sub-Artic sample**	p-value
Number of participants	N	18	20	
Gender	N (male/female)	10/8	1/19	0.000601
Age, *years*	Mean ± SD (Range)	42.8 ± 11.9 (20 - 61)	41.1 ± 13.2 (21 - 69)	0.685
Weight, *kg*	Mean ± SD	74.9 ± 13.0	73.4 ± 12.0	0.717
Height, *cm*	Mean ± SD	172.3 ± 10.5	167.0 ± 8.7	0.095
Body mass index, kg/m^2	Mean ± SD	25.1 ± 3.3	26.3 ± 3.9	0.319
Dominant hand	N (right-handed/left-handed)	16/2	18/2	
NRS, 0-100	Mean ± SD (Range)	15.0 ± 17.7 (0 - 40)	19.4 ± 11.1 (0 - 40)	0.373
Skin temperature, *°C*	Mean ± SD	32.3 ± 1.0	32.7 ± 1.5	0.441
Room temperature*, *°C*	Mean ± SD	22.5 ± 0.6	22.5 ± 0.7	0.978

Notes: Data are presented as mean ± standard deviation.
N = number of participants; SD = standard deviation; NRS = numerical rating scale 0-100 (0 = "no pain", 100 = "worst pain imaginable"). * Temperature in the examination room.

2

Table 2. QST z-score values of the total sample, separated by groups (mid-European sample, N=18, and sub-Arctic sample, N=20) being analyzed.

		Mid-European sample (N=18)	Sub-Artic sample (N=20)	p-value
CDT[a]	Mean ± SD (SEM) 95 % CI of mean	0.02 ± 0.98 (0.23) -0.47 to 0.50	0.37 ± 0.64 (0.14) 0.07 to 0.67	0.200
WDT[a]	Mean ± SD (SEM) 95 % CI of mean	0.07 ± 0.75 (0.18) -0.30 to 0.44	-0.28 ± 0.47 (0.10) -0.49 to -0.06	0.088
TSL[a]	Mean ± SD (SEM) 95 % CI of mean	-0.33 ± 0.72 (0.17) -0.69 to 0.03	-0.47 ± 0.56 (0.13) -0.73 to -0.20	0.509
CPT	Mean ± SD (SEM) 95 % CI of mean	-0.38 ± 0.71 (0.17) -0.74 to -0.03	0.64 ± 1.05 (0.23) 0.15 to 1.13	0.001297
HPT	Mean ± SD (SEM) 95 % CI of mean	-0.73 ± 0.72 (0.17) -1.09 to -0.37	0.19 ± 0.88 (0.19) -0.22 to 0.60	0.000189
PPT[a]	Mean ± SD (SEM) 95 % CI of mean	0.28 ± 0.93 (0.22) -0.18 to 0.75	0.13 ± 0.80 (0.18) -0.25 to 0.50	0.581
MPT[a]	Mean ± SD (SEM) 95 % CI of mean	0.10 ± 0.94 (0.22) -0.37 to 0.56	0.33 ± 0.92 (0.21) -0.10 to 0.76	0.447
MPS[a]	Mean ± SD (SEM) 95 % CI of mean	0.21 ± 0.81 (0.19) -0.19 to 0.61	0.10 ± 1.00 (0.22) -0.37 to 0.57	0.711
WUR[a]	Mean ± SD (SEM) 95 % CI of mean	0.01 ± 0.85 (0.20) -0.41 to 0.44	0.27 ± 0.70 (0.16) -0.06 to 0.59	0.325
MDT[a]	Mean ± SD (SEM) 95 % CI of mean	-0.67 ± 0.65 (0.15) -1.00 to -0.35	-0.49 ± 0.84 (0.19) -0.88 to -0.09	0.453
VDT	Mean ± SD (SEM) 95 % CI of mean	-1.09 ± 0.88 (0.21) -1.53 to -0.65	-1.32 ± 0.56 (0.13) -1.59 to -1.06	0.356

Notes: QST z-score values are presented as mean ± standard deviation (standard error of the mean) and 95 % confidence interval of mean.
[a]Normal distribution tested by using the Shapiro-Wilk test. Post-hoc analysis by using independent t-tests adjusted according to Bonferroni with a resulting p-value of $p \leq 0.0045$.
Abbreviations: QST = quantitative sensory testing; CDT = cold detection threshold; WDT = warm detection threshold; TSL = thermal sensory limen; CPT = cold pain threshold; HPT = heat pain threshold; PPT = pressure pain threshold; MPT = mechanical pain threshold; MPS = mechanical pain sensitivity; WUR = wind-up ratio; MDT = mechanical detection threshold; VDT = vibration detection threshold; SD = standard deviation; SEM = standard error of the mean; 95 % CI of mean = 95 % confidence interval of the mean.

Table 3. QST z-scores values of all females, separated by groups (mid-European females, N=8, and sub-Arctic females, N=19) being analyzed.

		Mid-European sample (N=8)	**Sub-Artic sample (N=19)**	p-value
CDT[a]	Mean ± SD (SEM) 95 % CI of mean	-0.41 ± 1.13 (0.40) -1.36 to 0.54	0.34 ± 0.64 (0.15) 0.03 to 0.65	0.113
WDT[a]	Mean ± SD (SEM) 95 % CI of mean	0.31 ± 0.79 (0.28) -0.35 to 0.96	-0.15 ± 0.45 (0.11) -0.37 to 0.07	0.066
TSL[a]	Mean ± SD (SEM) 95 % CI of mean	-0.20 ± 0.59 (0.21) -0.69 to 0.30	-0.46 ± 0.58 (0.13) -0.74 to -0.18	0.290
CPT	Mean ± SD (SEM) 95 % CI of mean	-0.31 ± 0.69 (0.24) -0.88 to 0.27	0.72 ± 1.01 (0.23) 0.23 to 1.21	0.015
HPT	Mean ± SD (SEM) 95 % CI of mean	-0.53 ± 0.86 (0.31) -1.25 to 0.19	0.35 ± 0.86 (0.20) -0.07 to 0.76	0.023
PPT[a]	Mean ± SD (SEM) 95 % CI of mean	0.57 ± 0.84 (0.30) -0.14 to 1.27	0.13 ± 0.82 (0.19) -0.26 to 0.53	0.226
MPT[a]	Mean ± SD (SEM) 95 % CI of mean	-0.20 ± 1.01 (0.36) -1.04 to 0.64	0.38 ± 0.91 (0.21) -0.06 to 0.82	0.157
MPS[a]	Mean ± SD (SEM) 95 % CI of mean	-0.01 ± 0.89 (0.32) -0.76 to 0.74	0.19 ± 0.93 (0.21) -0.25 to 0.64	0.608
WUR[a]	Mean ± SD (SEM) 95 % CI of mean	0.16 ± 0.90 (0.32) -0.59 to 0.92	0.28 ± 0.71 (0.16) -0.07 to 0.62	0.730
MDT[a]	Mean ± SD (SEM) 95 % CI of mean	-0.59 ± 0.47 (0.17) -0.98 to -0.20	-0.55 ± 0.81 (0.19) -0.94 to -0.16	0.869
VDT	Mean ± SD (SEM) 95 % CI of mean	-1.46 ± 0.50 (0.18) -1.88 to -1.04	-1.33 ± 0.58 (0.13) -1.61 to -1.05	0.574

Notes: QST z-scores values are presented as mean ± standard deviation (standard error of the mean) and 95 % confidence interval of the mean. [a]Normal distribution tested by using the Shapiro-Wilk test. Post-hoc analysis by using independent t-tests adjusted according to Bonferroni with a resulting p-value of p≤0.0045.

Abbreviations: QST = quantitative sensory testing; CDT= cold detection threshold; WDT = warm detection threshold; TSL = thermal sensory limen; CPT = cold pain threshold; HPT= heat pain threshold; PPT= pressure pain threshold; MPT = mechanical pain threshold; MPS = mechanical pain sensitivity; WUR = wind-up ratio; MDT = mechanical detection threshold; VDT= vibration detection threshold; SD = standard deviation; SEM = standard error of the mean; 95 % CI of mean = 95 % confidence interval of the mean.

Table 4. Pain intensity scores (NRS, 0-100) of the total sample, separated by groups (mid-European sample, N=18, and sub-Arctic sample, N=20) and separated by recruited as healthy control and pain patient being analyzed.

		Mid-European Sample		Sub-Artic sample		p-value
Number of participants	N	18		20		
NRS, 0-100	Mean ± SD	15.0 ± 17.7		19.4 ± 11.1		0.373
	(Range)	(0 - 40)		(0 - 40)		
		recruited as healthy control	**recruited as pain patient**	**recruited as healthy control**	**recruited as pain patient**	
Number of participants (separated in groups)	N	10	8	1	19	0.000601
NRS, 0-100 (separated in groups)	Mean ± SD		33.8 ± 6.5		20.4 ± 10.4	0.001
	(Range)		(20 - 40)		(5 - 40)	

Notes: Pain intensity scores with a numeric rating scale (NRS, ranging from 0-100) are presented as mean ± standard deviation (range) for mid-European and sub-Arctic sample, and when separated into the groups healthy control and pain patient.
Abbreviations: NRS, numerical rating scale; SD = standard deviation.

Table 5. Correlations of pain intensity scores (NRS, 0-100) and CPT z-score values and HPT z-score values of all participants recruited as pain patients (N=27) and separated by groups (mid-European sample, N=8), and sub-Arctic sample, N=19) being analyzed.

	Pain patients in the total sample (N=27)	Pain patients in the Mid-European sample (N=8)	Pain patients in the Sub-Arctic sample (N=19)
NRS x CPT	r=-0.049; p=0.722	r=0.148; p=0.615	r=0.149; p=0.380
NRS x HPT	r=-0.298; p=0.031	r=0.148; p=0.615	r=-0.077; p=0.648

Notes: Correlations between NRS and z-transformed CPT and HPT using by Kendall-Tau-b presented for all participants recruited as pain patients, and mid-European and sub-Arctic sample.
Abbreviations: NRS = numerical rating scale; CPT = cold pain threshold; HPT = heat pain threshold; r = correlation coefficient.

Table 6. ANCOVA of CPT z-scores values and HPT z-scores values with NRS (0-100) as covariant of all participants recruited as pain patients (N=27) being analyzed.

		Mid-European sample (N=8)	Sub-Arctic sample (N=19)		NRS as covariant
CPT	mean ± SD	-0.48 ± 0.63	0.72 ± 1.01	F(1,24) = 14.48; p=0.001; ε=0.376	F(1,24) = 3.98; p=0.058; ε=0.142
HPT	mean ± SD	-0.98 ± 0.58	0.35 ± 0.86	F(1,24) = 9.97; p=0.004; ε=294	F(1,24) = 0.03; p=0.866; ε=0.001

Notes: ANCOVA of CPT z-scores values and HPT z-scores values with pain intensity (NRS 0-100) as covariant are presented as F- and p-values, and effect size of all participants recruited as pain patients (N=27).
Abbreviations: CPT = cold pain threshold; HPT = heat pain threshold; NRS = numerical rating scale; SD = standard deviation.

Table 7. Comparison of QST z-scores values between two examiners (HN and BD) of the sub-Arctic sample (N=20) being analyzed.

		measured by HN (N=13)	**measured by BD (N=7)**	p-value
CDT	Mean ± SD (SEM)	0.47 ± 0.71 (0.20)	0.20 ± 0.48 (0.18)	0.384
WDT	Mean ± SD (SEM)	-0.12 ± 0.47 (0.13)	-0.30 ± 0.43 (0.16)	0.417
TSL	Mean ± SD (SEM)	-0.41 ± 0.55 (0.15)	-0.57 ± 0.61 (0.23)	0.550
CPT	Mean ± SD (SEM)	0.53 ± 1.14 (0.32)	0.85 ± 0.88 (0.33)	0.527
HPT	Mean ± SD (SEM)	0.17 ± 0.92 (0.25)	0.63 ± 0.60 (0.23)	0.249
PPT	Mean ± SD (SEM)	0.10 ± 0.78 (0.22)	0.17 ± 0.89 (0.34)	0.853
MPT	Mean ± SD (SEM)	0.24 ± 0.81 (0.23)	0.50 ± 1.14 (0.43)	0.566
MPS	Mean ± SD (SEM)	0.06 ± 1.17 (0.33)	0.17 ± 0.63 (0.24)	0.781
WUR	Mean ± SD (SEM)	0.24 ± 0.71 (0.20)	0.32 ± 0.73 (0.27)	0.803
MDT	Mean ± SD (SEM)	-0.52 ± 0.93 (0.26)	-0.43 ± 0.69 (0.26)	0.830
VDT	Mean ± SD (SEM)	-1.27 ± 0.63 (0.18)	-1.42 ± 0.44 (0.16)	0.574

Notes: QST z-scores values are presented as mean ± standard deviation (standard error of the mean (SEM) of all participants of sub-Arctic sample (N=20).
Abbreviations: QST = quantitative sensory testing; CDT = cold detection threshold; WDT = warm detection threshold; TSL = thermal sensory limen; CPT = cold pain threshold; HPT = heat pain threshold; PPT = pressure pain threshold; MPT = mechanical pain threshold; MPS = mechanical pain sensitivity; WUR = wind-up ratio; MDT = mechanical detection threshold; VDT = vibration detection threshold; SD= standard deviation; SEM= standard error of the mean.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, Helen Nothnagel, dass mir die geltende Promotionsordnung der Fakultät für Sozial- und Verhaltenswissenschaften der Friedrich-Schiller-Universität Jena bekannt ist. Ich habe die Dissertation selbst angefertigt. Alle von mir benutzten Hilfsmittel, verwendeten Quellen und persönlichen Mitteilungen habe ich in meiner Arbeit angegeben. Alle Personen, die bei der Auswahl und der Auswertung des Materials beteiligt waren, sind namentlich in der Dissertation unter dem Abschnitt *Angaben zum Eigenanteil* aufgeführt. Ich erkläre weiterhin, dass ich keine Hilfe eines Promotionsberaters in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Dissertation wurde noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Zudem wurde keine gleiche, keine in wesentlichen Teilen ähnliche und auch keine andere Abhandlung bei einer anderen Hochschule bzw. anderen Fakultät als Dissertation eingereicht.

Ich versichere, nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen zu haben.

Jena, den 29. März 2021 Helen Nothnagel