PAIN® 151 (2010) 598–605

www.elsevier.com/locate/pain

Reference data for quantitative sensory testing (QST): Refined stratification for age and a novel method for statistical comparison of group data

Walter Magerl [a,*,1], Elena K. Krumova [b,1], Ralf Baron [c], Thomas Tölle [d], Rolf-Detlef Treede [a], Christoph Maier [b]

[a] Department of Neurophysiology, Center for Biomedicine and Medical Technology Mannheim (CBTM), Medical Faculty Mannheim, Ruprecht Karls University Heidelberg, Ludolf-Krehl-Strasse 13-17, 68167 Mannheim, Germany
[b] Department of Pain Management, BG Kliniken Bergmannsheil, Ruhr-University of Bochum, Bürkle-de-la-Camp-Platz 1, 44789 Bochum, Germany
[c] Division of Neurological Pain Research and Therapy, Department of Neurology, University of Kiel, Germany
[d] Department of Neurology, Technische Universität München, Germany

ARTICLE INFO

Article history:
Received 6 October 2009
Received in revised form 5 July 2010
Accepted 22 July 2010

Keywords:
Quantitative sensory testing
Reference data
Stratification
Statistics
Gender
Age

ABSTRACT

Clinical use of quantitative sensory testing (QST) requires standardization. The German research network on neuropathic pain (DFNS) solves this problem by defining reference data stratified for test site, gender and age for a standardized QST protocol. In this report we have targeted two further problems: how to adjust for age-related sensory changes, and how to compare groups of patients with the reference database. We applied a moving average across ages to define reference values per decade. This analysis revealed that women were more sensitive to heat pain independent of age. In contrast, functions were converging at older age for blunt pressure pain, but diverging for punctate mechanical pain (pin prick). The probability that an individual patient dataset is within the range of normal variability is calculated by z-transform using site-, gender- and age-specific reference data. To compare groups of patients with reference data, we evaluated two techniques: A: paired *t*-test versus fixed mean; i.e. the reference mean value is considered as the known population mean, B: non-paired *t*-test versus the reference dataset and number of cases restrained to the same number of cases as the patient data set. Simulations for various sample sizes and variances showed that method B was more conservative than method A. We present a simple way of calculating method B for data that have been z-normalized. This technique makes the DFNS reference data bank applicable for researchers beyond the DFNS community without a need for subsampling of subjects from the database.

1. Introduction

The German Research Network on Neuropathic Pain (DFNS) has previously introduced a standardized protocol for quantitative sensory testing (QST) in humans [27]. The DFNS approach to obtain a comprehensive profile of somatosensory functions within a reasonably short period of time was also considered useful by researchers outside the DFNS [5,23]. Reference data in a multi-center cohort of healthy subjects of both genders were based on this QST protocol [28]. Consistent with prior studies (e.g. [3,9,14,26,32], these reference values were dependent on test site, gender and age. The DFNS proposed to normalize individual patient data to group-specific means and standard deviations by *z*-transform and to consider *z*-values below −1.96 or above +1.96 as abnormal for diagnostic purposes (95% confidence interval [33]).

In the DFNS reference data, age-related differences in cutaneous sensitivity were estimated by dividing the cohort into young (<40 years) and old subjects (⩾40 years). Whereas this approach was appropriate to demonstrate the presence of age-related differences for most QST-parameters (with the exception of pin prick-evoked measures MPT: mechanical pain threshold, MPS: mechanical pain sensitivity, and WUR: wind-up ratio to pin prick), the dichotomy created a discontinuity for the *z*-transform when crossing the age of 40 years. This discontinuity will be misleading in longitudinal studies, since a mildly abnormal QST value for age 39 falls into the normal range for age 41. In this paper we report decade-specific reference values generated by a moving-average technique that minimizes discontinuities. We then use these values to re-assess gender differences as a function of age.

Abbreviations: CDT, cold detection threshold; CPT, cold pain threshold; DFNS, Deutscher Forschungsverbund Neuropathischer Schmerz (= German Research Network on Neuropathic Pain); DMA, dynamic mechanical allodynia; HPT, heat pain threshold; MDT, mechanical detection threshold; MPS, mechanical pain sensitivity; MPT, mechanical pain threshold; PHS, paradoxical heat sensation; PPT, pressure pain threshold; QST, quantitative sensory testing; TSL, thermal sensory limen; VDT, vibration detection threshold; WDT, warm detection threshold; WUR, wind-up ratio.

* Corresponding author.
E-mail address: walter.magerl@medma.uni-heidelberg.de (W. Magerl).

[1] Both authors contributed equally to this study.

0304-3959/$36.00
doi:10.1016/j.pain.2010.07.026

The site-, gender- and age-specific reference data in this paper may be used by any laboratory to compare their data with a reference dataset of 1080 values. The validity of such a comparison depends on using the same QST testing protocol, for which formal training and certification have been established [12]. Whereas individual patient data are easily evaluated via the *z*-transform and the 95% confidence interval, no procedure has been identified yet for the statistical comparison of group data with this reference data set. Such a procedure is urgently needed since already 117 investigators from 38 groups in 15 different countries have been trained by certified DFNS centers and are using the DFNS' QST system for basic and clinical science, in academia as well as in the pharmaceutical industry. This paper provides a simple algorithm for group comparisons that does not require access to the reference dataset itself and establishes rules for statistical comparison. Moreover, it suggests quality criteria for centers for comparison of their own data to the DFNS reference data.

2. Methods

To obtain an age-adjusted reference data base, the recently published reference data [28] were re-analyzed and separated for both genders and the three body regions that were available in the DFNS data bank: face (mid cheek, blunt pressure pain threshold PPT on the masseter muscle with upper and lower teeth in occlusion position), hands (dorsum, blunt pressure pain threshold PPT at the thenar) and feet (dorsum, blunt pressure pain threshold PPT at the instep). Currently we do not have multi-center reference data for other body regions yet. We tentatively use hand data as representative for the upper body and foot data for the lower body. Unpublished single-center data indicate that in spite of mean value differences the 95% confidence intervals overlap vastly for most QST parameters for measurements on the dorsal hand compared to measurements on the palmar hand, in the thoracic areas or in the areas over the trapezius muscle (except for higher mechanical and thermal detection thresholds on the palmar side and for higher vibration detection thresholds on the trunk). Since left–right differences were independent of test region [28], it is prudent to rely more on those differences than on absolute reference values when evaluating non-standard test regions.

The data were obtained from 180 healthy subjects in a comprehensive standardized QST protocol consisting of 7 tests measuring 13 parameters. All subjects who were entered in this data set were of European/Caucasian descent. Data from both body sides were pooled because all correlations across the two body sides were highly significant (all $p < 0.001$), regression functions were close to unity and there was no significant difference. For details of the protocol, see parent data set [28]. Briefly, the protocol contained both thermal and mechanical test stimuli, namely: thermal detection thresholds for the perception of cold (CDT: cold detection threshold), warmth (WDT: warm detection threshold) and paradoxical heat sensations (PHS: paradoxical heat sensation), thermal pain thresholds for cold (CPT) and hot stimuli (HPT), mechanical detection thresholds for touch (MDT) and vibration (VDT), mechanical pain sensitivity including thresholds for pinprick (MPT) and blunt pressure (PPT), a stimulus–response-function for pinprick sensitivity (MPS) and dynamic mechanical allodynia (DMA: dynamical mechanical allodynia) as well as pain summation to repetitive pinprick stimuli (WUR¸ wind-up ratio) [27].

The data base was divided according to age into five segments representing decade groups. To avoid discontinuities, a moving-average procedure was used by including the neighboring half-decades for calculation of mean and standard deviation of each age group:

- decade 20–30 years: calculated from subjects between > 15 and 35 years of age
- decade 30–40 years: calculated from subjects between > 25 and 45 years of age
- decade 40–50 years: calculated from subjects between > 35 and 55 years of age
- decade 50–60 years: calculated from subjects between > 45 and 65 years of age
- decade 60–70 years: calculated from subjects between > 55 and 75 years of age

All data except for PHS, CPT, HPT and VDT were normally distributed in logarithmic space (log-normal distribution) as already shown for this data set in previous analysis [28]. Thus, logarithmical transformation was performed for all other parameters prior to statistical analysis to achieve secondary normal distribution (for theoretical background, see also [27]. Correlation analysis between the subjects' age and QST parameters was performed using bivariate parametric correlation/regression analysis (Pearson). Differences between age decade, gender (between-subjects factors) and tested body region (within-subject factor) were compared using three-way mixed model ANOVA for all QST parameters.

Differences between the ten DFNS centers contributing subjects to QST reference data were analyzed to delineate the magnitude of between-centers variation, and to derive measures of distribution (standard deviation, confidence intervals) within and across centers. Furthermore, effect sizes for every center were calculated to judge the magnitude of deviation from the grand mean. These distribution parameters are used to delineate guidelines for comparative self-checking of centers outside of DFNS.

A novel method of statistical comparison was developed, which allows a bias-free and balanced comparison of data sets of any mixture of QST data based on the DFNS battery of sensory assessment. Two different approaches were compared using calculation of *t*-tests on standard normal data (*z*-values) for patient groups with a wide variety of means, standard deviations and numbers of subjects to illustrate the feasibility and statistical merits of the methods.

- In the first approach, the patient groups were compared to the reference data by treating the reference data mean ($z = 0$) as the known population mean value. This is a standard procedure, which is implemented in many statistical software packages and is equivalent to paired *t*-tests versus an equal number of zeros.
- In the second approach, the patient groups were compared to the reference data by non-paired *t*-tests. Such an approach may lead to false positive findings due to the large number of degrees of freedom from the reference data (1080 values). Therefore, we introduced a virtual subsampling of the reference data base, by setting its "*n*" number equal to the number of observations in the patient data set.
- Data in figures are presented as mean ± SD. Reference data are shown as means and 95% confidence intervals (mean ± 1.96 * SD). Data of log-transformed QST parameters (CDT, WDT, TSL, MDT, MPT, MPS, ALL, WUR and PPT) were retransformed to values representing the original units of the parameters.

3. Results

3.1. New age-related QST reference data

The reference data published in [28] have been re-analyzed with a higher resolution for age (Table S1). Analysis of variance confirmed that non-nociceptive thermal (CDT, WDT, TSL) and tactile thresholds (MDT) varied by region always in the same rank

order (feet > hand > face), but did essentially not differ between genders (Table 1). We now found that these non-nociceptive thresholds varied significantly linearly with age albeit correlation coefficients were very small (correlations between $r = 0.207$ and $r = 0.257$, all $p < 0.01$), and the percentage of variance explained by the covariate age (approximately 4–6%) was negligible (see Table S2). Generally, thresholds increased by approximately 50% across the age range tested (age dependency was somewhat lesser in the face).

As previously shown, nociceptive thresholds exhibited significant gender differences, females being more sensitive than males (Table S1). The refined stratification for age enabled a more detailed analysis of the interaction of age and gender. Little complexity was found in the analysis of thermal pain thresholds. Heat pain thresholds increased monotonically with age in all test regions, being approximately 2.1 °C higher in the oldest cohort (60–70 years) than in young adults (20–30 years). This threshold increase was independent of gender, and heat pain thresholds were 1.6 °C higher in male than in female subjects throughout all age ranges (Fig. 1). Cold pain thresholds of male subjects were on average met at approximately 1.7 °C lower temperatures than in female subjects (Fig. 2A). In females, cold pain thresholds dropped monotonically with increasing age by approximately 9 °C, while age variation was lesser and also less consistent in male subjects, which is substantiated as a significant age × gender interaction (ANOVA, $p < 0.01$, Table 1).

Gender differences across age ranges followed a more complex pattern for mechanical pain thresholds (Fig. 2B and C). Differences between genders systematically shrank for blunt pressure (PPT), and diverged for pin pricks (MPT) with increasing age. Pain thresholds to blunt pressure stimuli (PPT) differed between male and female subjects in young and middle-aged adults (20–50 years, all $p < 0.005$), but PPT did not differ any more between male and female subjects in the older cohorts (50–70 years, $p > 0.60$, each). Conspicuously, MPTs were the only thresholds with no overall dependency on age (overall correlation with age: $r = -0.067$, n.s.). Nevertheless, thresholds in male and female varied with age, but they did so in opposite ways in all body regions (depicted for MPT at the hand in Fig. 2C). While MPTs were not different at all between genders at young age (post hoc significances for the lower two decades: $p = 0.47$ and $p = 0.77$), thresholds of male and female subjects progressively diverged symmetrically with increasing age. Accordingly, MPT in older males was approximately twice as high as in females (post hoc significances for the upper three decades: $p < 0.005$, $p < 0.02$, and $p < 0.01$).

3.2. Comparing patient group data to the QST reference dataset

Patient groups are typically inhomogeneous with respect to gender and age, and test sites may also differ (e.g. peripheral nerve injury of upper and lower limb). Such data may nonetheless be averaged, provided each value is first normalized to the appropriate subgroup in the stratified reference dataset (Appendix 1) by first subtracting the subgroup-specific mean and then dividing by the respective standard deviation (z-transform). If sensory function in the patient group is unaffected, their data distribution can be expected to have zero mean and standard deviation of one. An intuitive approach to test this null hypothesis would be by a non-paired t-test, according to the following equation:

$$t = (\text{mean}_{pat} - \text{mean}_{ref})/\text{square root}(\text{SD}^2_{pat}/n_{pat} + \text{SD}^2_{ref}/n_{ref}) \text{ and } df = n_{pat} + n_{ref} - 2 \quad (1)$$

Since the patient groups are typically at least one order of magnitude smaller than the reference population ($n_{pat} << n_{ref}$), this

Table 1
Correlation with age (percent of variance explained by age) and analysis of variance of QST reference data.

QST	Correlation with age[a] (variance explained, %)	Gender (1)	Age (2)	Region (3)	1 x 2	1 x 3	2 x 3	1 x 2 x 3
CDT	0.210** (4.41%)	$F_{1,350} = 1.72^{n.s}$	$F_{4,350} = 9.00^{***}$	$F_{2,700} = 341.40^{***}$	$F_{4,350} = 0.72^{n.s.}$	$F_{2,700} = 13.35^{***}$	$F_{8,700} = 2.78^{**}$	$F_{8,700} = 1.58^{n.s}$
WDT	0.211** (4.45%)	$F_{1,350} = 4.66^{*}$	$F_{4,350} = 8.86^{***}$	$F_{2,700} = 529.47^{***}$	$F_{4,350} = 1.44^{n.s.}$	$F_{2,700} = 9.42^{***}$	$F_{8,700} = 3.55^{***}$	$F_{8,700} = 1.00^{n.s.}$
TSL	0.240** (5.76%)	$F_{1,350} = 0.66^{n.s.}$	$F_{4,350} = 7.81^{***}$	$F_{2,700} = 668.59^{***}$	$F_{4,350} = 1.05^{n.s.}$	$F_{2,700} = 10.94^{***}$	$F_{8,700} = 3.55^{***}$	$F_{8,700} = 2.62^{**}$
CPT	-0.209** (4.37%)	$F_{1,350} = 3.29^{(*)}$	$F_{4,350} = 8.50^{***}$	$F_{2,700} = 15.43^{***}$	$F_{4,350} = 2.49^{*}$	$F_{2,700} = 0.99^{n.s.}$	$F_{8,700} = 1.14^{n.s.}$	$F_{8,700} = 0.88^{n.s.}$
HPT	0.257*** (6.60%)	$F_{1,350} = 23.32^{***}$	$F_{4,350} = 7.83^{***}$	$F_{2,700} = 53.38^{***}$	$F_{4,350} = 1.93^{n.s.}$	$F_{2,700} = 0.22^{n.s.}$	$F_{8,700} = 2.52^{*}$	$F_{8,700} = 1.05^{n.s.}$
PPT	0.207** (4.28%)	$F_{1,348} = 17.63^{***}$	$F_{4,348} = 8.71^{***}$	$F_{2,696} = 786.23^{***}$	$F_{4,348} = 2.69^{*}$	$F_{2,696} = 0.83^{n.s.}$	$F_{8,696} = 4.07^{***}$	$F_{8,696} = 1.12^{n.s.}$
MPT	-0.067$^{n.s}$ (0.45%)	$F_{1,350} = 17.53^{***}$	$F_{4,350} = 2.19^{(*)}$	$F_{2,700} = 141.81^{***}$	$F_{4,350} = 2.30^{(*)}$	$F_{2,700} = 5.68^{**}$	$F_{8,700} = 3.68^{***}$	$F_{8,700} = 2.42^{*}$
MPS	0.001$^{n.s}$ (0.00%)	$F_{1,350} = 3.48^{(*)}$	$F_{4,350} = 1.07^{n.s.}$	$F_{2,700} = 25.16^{***}$	$F_{4,350} = 0.98^{n.s.}$	$F_{2,700} = 0.77^{n.s.}$	$F_{8,700} = 1.37^{n.s.}$	$F_{8,700} = 1.38^{n.s.}$
WUR	0.056$^{n.s.}$ (0.31%)	$F_{1,340} = 0.77^{n.s.}$	$F_{4,340} = 1.20^{n.s.}$	$F_{2,680} = 16.97^{***}$	$F_{4,340} = 2.17^{(*)}$	$F_{2,680} = 4.47^{*}$	$F_{8,680} = 1.93^{(*)}$	$F_{8,680} = 2.11^{*}$
MDT	0.207** (4.28%)	$F_{1,350} = 4.09^{*}$	$F_{4,350} = 8.07^{***}$	$F_{2,700} = 599.17^{***}$	$F_{4,350} = 1.23^{n.s.}$	$F_{2,700} = 12.51^{***}$	$F_{8,700} = 6.36^{***}$	$F_{8,700} = 1.54^{n.s}$
VDT	-0.248*** (6.15%)	$F_{1,348} = 1.01^{n.s.}$	$F_{4,348} = 11.47^{***}$	$F_{2,696} = 86.54^{***}$	$F_{4,348} = 0.56^{n.s.}$	$F_{2,696} = 3.45^{*}$	$F_{8,696} = 8.41^{***}$	$F_{8,696} = 2.46^{*}$

n.s. not significant
(*) $p < 0.10$
* $p < 0.05$
** $p < 0.01$
*** $p < 0.001$
[a] average correlation coefficients across genders and body regions.

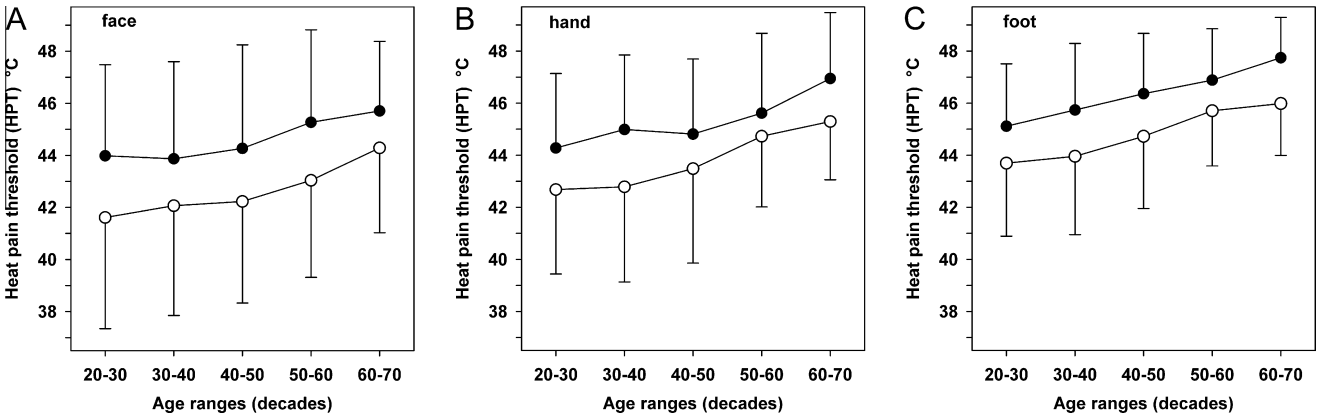


Fig. 1. Age dependence of heat pain threshold (HPT) in the face (A), hand (B) and foot (C). HPT increased monotonically with age in all regions, and was approximately 2.1 °C higher in the oldest cohort (50–60 years) than in young adults (20–30 years) independent of gender and test region. In all test areas, HPT in female subjects (open circles) was about 1.6 °C lower than in male subjects (solid circles) independent of age and test region. HPT was lowest in the face, intermediate on the hand and highest on the foot.

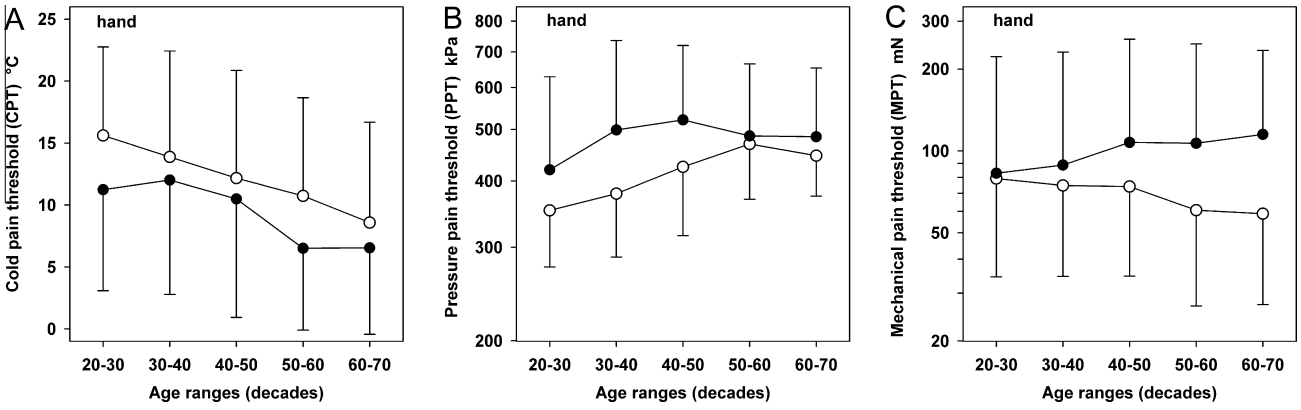


Fig. 2. Age dependence of pain thresholds in the hand for cold pain (CPT), blunt pressure pain (PPT) and pain to pin prick stimuli (MPT). (A) CPT shifted linearly to lower temperatures with age by approximately 7 °C from the youngest to the oldest cohort, and CPT in male subjects was met on average at 1.7 °C lower temperatures than in females with a nearly parallel shift between genders, but less regular in male (solid circles) than in female subjects (open circles). (B) PPT differed between male and female in young adults. PPT increased with age to level off at mid-age levels, which occurred earlier in males than in females resulting in convergence of PPT in the older cohorts (50–70 years). (C) MPT did not differ between genders at young age, but thresholds of male and female subjects progressively diverged with increasing age.

approach causes two problems: the denominator is dominated by the standard deviation of the patient group, which is estimated rather imprecisely with a small number of observations. Moreover, the degrees of freedom are dominated by the large reference group. These effects may lead to false positive results, in particular for small homogenous patient groups (small n and small SD; see also below).

As a first solution to this problem, we considered the standard statistical procedure to compare group data with the theoretical or known population mean by a paired t-test (method A). This solution is reasonable, since the population mean will be estimated well if the reference group is sufficiently large. Inflated degrees of freedom are avoided, since for paired t-tests only the number of observations in the patient group counts. However, the problem of inaccuracies in estimating the denominator remains, as illustrated in the following equation:

$$t = (\text{mean}_{\text{pat}} - 0)/\text{square root}(\text{SD}^2_{\text{pat}}/n_{\text{pat}}) \text{ and } df = n_{\text{pat}} - 1 \quad (2)$$

As an alternative solution to the problems with Eq. (1), we considered a virtual subsampling of the reference dataset (method B) such that mean and standard deviation are maintained, but its number of observations equals that in the patient group ($n_{\text{pat}} = n_{\text{ref}}$). Since the reference data have zero mean and unit standard deviation, this leads to a rather simple equation:

$$t = (\text{mean}_{\text{pat}} - 0)/\text{square root}(\text{SD}^2_{\text{pat}}/n_{\text{pat}} + 1/n_{pat}) \text{ and } df = 2 * n_{\text{pat}} - 2 \quad (3)$$

The relevance of reducing the number of subjects in the virtual control group is illustrated by the following example: Given an estimate of 0.3 ± 0.5 ($n = 32$) for the patient group and standard normal distribution for the control group (0 ± 1), this leads to a test statistic of $t = 3.21$ ($p = 0.003$), when tested against the full-size control group. However, when tested against a control group of the same size ($n = 32$), the test statistic becomes $t = 1.52$ ($p = 0.13$).

When comparing Eqs. (2) and (3), it is not intuitively evident as which one and under what circumstances will be more conservative: due to the additional term in its denominator the t-value in Eq. (3) is smaller than in formula 2, but it is associated with a larger number of degrees of freedom. We therefore performed simulations for a range of numbers of patients (10–100), standard deviations in the patient data smaller or larger than in the reference group (0.25–4.0), and mean values of the patient group between 0.25 and 1.0 (corresponding to effect sizes of 0.25–1.0).

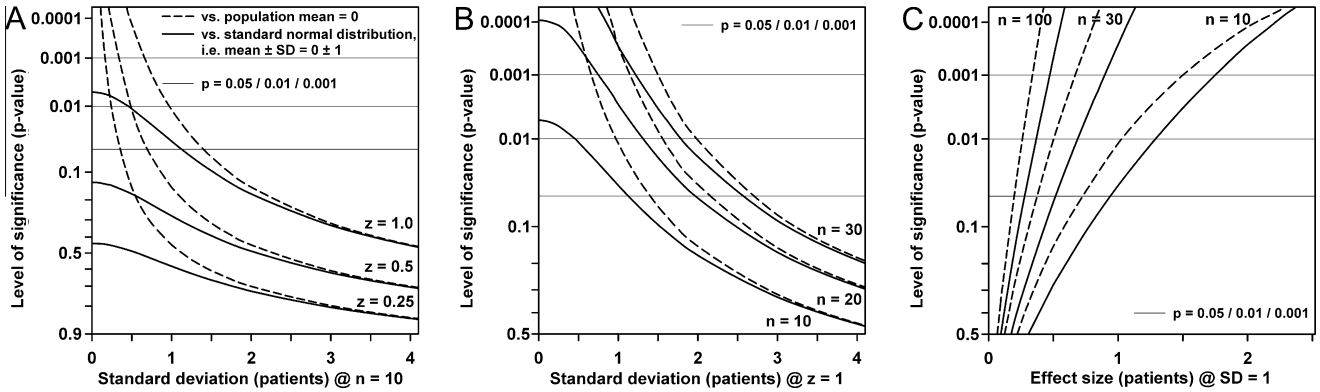


Fig. 3. Simulated comparisons of patient groups with QST reference dataset by two methods using *z*-transformed data. Dashed lines: Method A (paired *t*-test versus known population mean). Solid lines: Method B (non-paired *t*-test versus virtual subsample of the reference dataset yielding the same sample size as the patient dataset). (A) Estimation of *p*-values was more conservative for method B than for method A independent of effect size (indicated by *z*-value of the patient dataset), especially at lower standard deviations of the patient data set (for $n = 10$). (B) Estimation of *p*-values was also more conservative for method B than for method A independent of sample size, especially at lower standard deviations of the patient data set (for effect size = 1.0). (C) When plotted against effect size, the probability curves for method B were shifted to the right of method A, indicating that this more conservative method requires either a larger effect size or a larger sample size to be significant (curves calculated for standard deviation in patient group = 1.0).

Fig. 3A displays probability values for both scenarios for varying standard deviation of a small group of 10 patients. Probability curves converge for larger standard deviations, but for small standard deviations, method A (Eq. (2)) over-estimates the *t*-value leading to higher levels of significance, independent of effect sizes (equal to the mean *z*-value of the patient group). Fig. 3B displays probability values for varying standard deviations at a fixed effect size of 1.0. Also for larger group sizes, method A (Eq. (2)) led to less conservative estimates than method B (Eq. (3)). In Fig. 3C, *p*-values are plotted versus effect size. Method B (solid lines) usually requires about 30% larger effect size than method A (dashed lines) to reach the same level of significance. Thus, for any combination of sample size, effect size and standard deviation in the patient group, more conservative estimates were obtained by method B that takes the standard deviation of the reference dataset into account. Probability curves obtained by methods A and B converged only at very large standard deviations of the patient group (i.e. $SD_{pat} >> SD_{ref}$ – Fig. 3A and B) or at (unrealistically) high effect sizes (Fig. 3C).

Finally, we have analyzed differences between the ten DFNS centers contributing subjects to QST reference data. Although between-center data were not strictly designed to analyze such differences, this analysis may provide some guideline for other (non-DFNS) centers, who may want to compare their data to the DFNS data. To circumvent sampling asymmetries between centers this analysis used *z*-transformed values, which cancel the effects of area, age and gender. The results of all centers are listed in Table S2. As expected, the mean *z*-value ± SD was almost 0 ± 1 (a small deviation to 0.01 ± 0.99 was caused by few missing single data). The mean of *z*-values of the single centers ranged from −0.21 to +0.18 with a grand mean and 95% confidence interval of 0.01 ± 0.25. Likewise, standard deviations of the centers varied between 0.93 and 1.09 with a grand mean and 95% confidence interval of 0.99 ± 0.10. We suggest that any laboratory wishing to use the DFNS reference data should have mean and SD of their local z-transformed data from healthy subjects within these 95% confidence intervals.

4. Discussion

The German Research Network on Neuropathic Pain (DFNS) has implemented a multi-center database of QST reference values for both genders and currently three body regions, namely face, hand and foot [28]. Transformation into standard normal distribution, i.e. *z*-transformation, allows an easy judgment, whether a value in a given patient is outside the normal range defined as 95% confidence interval (±1.96 * SD; note that several parameters need to be logarithmically transformed for this purpose [27]).

Whereas gender and test site can be regarded as discrete variables reasonably taken into account by a stratified reference dataset [3,16,20] (but see [15] for masculinity/femininity as a potential continuous variable), age clearly is a continuous variable. Stratification into two age groups as in the DFNS data base and in other publications [4,6,7,9] may lead to age-related bias if used diagnostically. For example, average heat pain thresholds in the foot were 45.1 versus 47.0 °C in subjects of the young vs. old age cohort [28]. Thus, the same QST result in two patients of similar age (e.g. 38 and 42 years) would be interpreted differently, e.g. a heat pain threshold of 42 °C in the foot may be diagnosed as normal in the 38 year old (nominally "young") subject, but as abnormal (hyperalgesic) in the 42 year old (nominally "old") subject. Conversely, younger and older patients (e.g. 42 and 67 years) of the same age cohort may be misdiagnosed in the opposite direction.

4.1. The regression approach

Although there were highly significant correlations with age throughout the majority of QST parameters, continuous adjustment for age of the subject/patient is problematic due to the low value of correlation coefficients. Similarly low correlation with age has also been reported for nociceptive and non-nociceptive QST thresholds in other large scale studies (e.g. [21,32]) and for amplitudes of somatosensory-evoked potentials, an objective measure of somatosensory function [39]. When the correlation coefficient is low, the slope of the regression line is low too. A correction of age-dependence by regression would return a tilted regression function much shallower than real (e.g. as estimated by eye-fit) resulting in an age-related evaluation bias and the function would predict unrealistic values in young (higher) and old subjects (lower) and a relative insensitivity to detect sensory loss in young subjects, but overestimate the loss in old subjects. In contrast, the opposite would be true for sensory gain, thus overestimating hyperalgesia symptoms in young, but underestimating them in old subjects. Moreover, age-dependence was not linear for all parameters (e.g. for PPT), and it varied between genders for the same parameter (e.g. for MPT). Moreover the explanatory power of this regression (e.g. when entered as a covariate to single out age-related variance) is very low,

since it covers only about 4–6% of total variance as shown in Table 1. Thus, the regression approach is of little value for QST data.

4.2. The reference cohort approach

Adequate norms for age require a better resolution than previously offered in the DFNS database. The larger number of age groups, however, diminishes the number of subjects per group, and hence the estimates of means and standard deviations become less accurate. Moreover, the discontinuities at the boundaries between age groups remain. The moving-average filter is a standard tool in digital signal processing when smoother functions are desired [2]. Its application ranges from EEG analysis to predictive epidemiology [8,35]. In this study, we increased the number of age groups from 2 to 5 and included half of each neighboring decade in a moving-average procedure for parameter estimation (e.g. for the decade 30–40 years, data from subjects between 25 and 45 years were used). This way, we were able to distinguish different types of age dependence of gender differences: parallel shift independent of age (HPT/CPT), convergence with age (PPT) and divergence with age (MPT). Such a differential description will be important e.g. when gender differences are related to differences in hormonal status that varies as a function of age [1,15].

Significant gender differences in heat and cold pain thresholds (HPT/CPT) regardless of age are consistent with previous studies [reviewed in [10,13,16,25]]. Gender difference at young age and threshold convergence at old age for blunt pressure pain thresholds (PPT) have been reported previously [24] suggesting disappearance with menopause. This is consistent with absence of gender difference in children and threshold divergence emerging at puberty [1]. For pin prick sensitivity previous studies did not find gender differences [1,29,30]. In our study differences became only significant at ages > 40. Two previous studies in seniors [37,38] did not report gender, thus gender and age interaction (progressively diverging towards older age) is a new finding.

4.3. Statistical approach to use the reference data base for group comparisons

We developed a novel method to compare data of experimental or patient groups statistically with the DFNS reference data set. The use of the whole set of reference data was discarded on grounds of inflated degrees of freedom, a serious violation of "fair" statistical comparison, since any case-control comparison is based on the implicit assumption of equal group sizes in experimental and control group [11,19,31]. An alternative strategy providing a representative source of control subjects, by narrowing the data bank pool via subsampling to the appropriate group size of the patient group was also discarded. Drawing a matched subsample from the database would produce a suitable cohort of control subjects. Using case-matching software based on statistical selection criteria by e.g. multidimensional scaling to identify nearest matches in age, gender, weight, height, blood pressure, body mass index etc., will return a perfect twin control subject creating a yoked control design [18]. However, such an approach requires access to the primary data in the data bank and is quite labor intensive.

Instead, we decided to use a virtual subsample by maintaining the estimates of mean and standard deviation from the entire reference dataset, but arbitrarily decreasing its sample size parameter to the sample size of the patient group (method B). We compared that approach to a standard test, paired t-test versus a known population mean (method A). Simulations shown in Fig. 3 demonstrate that method B is more conservative particularly when the standard deviation in the patient group is smaller than in the reference group. This outcome is plausible, since only method B and not method A takes the standard deviation of the reference group into account. However, method B is more conservative than method A also for small sample sizes down to $n = 10$, which was unexpected since method B uses a larger number of degrees of freedom in its t-test. Thus, method B was more conservative for all conditions considered, making it the method of choice for group comparisons.

The virtual control group approach avoids cumbersome selection procedures and obeys the principle of equal group size [11,19,31]. The virtual control group is characterized by a mean ± SD of 0 ± 1 with the same number as in the patient group. It does not necessitate access to the data bank, and will, thus, be accessible to anyone beyond the inner circle of DFNS members. Calculations can be run by using a probability calculator, a simple software readily available as internet freeware (e.g. the award-winning web-based software SISA [34], Appendix 1). Finally, the statistical strategy presented in this paper will also make the DFNS reference data base available for scientists beyond the DFNS community.

Whether or not a center may be eligible for such comparison necessitates formal criteria. Such quality criteria for other (non-DFNS) centers can be developed from analysis of variation between contributing DFNS centers. As a note of caution, these data were not strictly designed to analyze such differences, since sample size in each of the centers was small (18 subjects × 3 body areas × 2 body sides = 108 assessments, each). Moreover, sampling was not population-based and only controlled for age (young vs. old), but not balanced for gender or gender × age combination [28]. Analysis of center data revealed that the deviation of any single center from the grand mean ranged between −0.21 and +0.18 z-values, and effect sizes were always smaller than 0.20, which according to conventional classification [17,22,36] was very small. We suggest that any (non-DFNS) center wishing to compare their data to the DFNS reference data should assess a sample of comparable size, i.e. approximately 100 independent test areas in healthy subjects, which may be done by any combination of age, gender and test areas (currently limited to face, hand, and foot). These data may be transformed into standard normal data (z-values) using the data supplied in Table S1 of this paper and mean ± SD calculated across all z-transformed values of the whole data set. We suggest that the calculated difference should be within 95% confidence intervals of the between-center analysis of the DFNS given in Table S2, i.e. a mean difference <0.25 z-values and a SD within 1 ± 0.1. All participating centers of the DFNS were found to be within these confines. For an even more rigorous method of center validation, the DFNS has established a formal QST certification process [12].

Conflicts of interest

The authors declare that there were no conflicts of interest.

Acknowledgement

This work was supported by BMBF (German Ministry of Education and Research) grants to DFNS (German Research Network on Neuropathic Pain, 01EM0506). In the discussion of statistical analysis we received most helpful support by Daan Uitenbroek (Quantitative Skills and Municipal Health Service, Amsterdam, Netherlands). Skillful support for art work was provided by Martin Dettling.

Appendix A. Comparison of group data to reference data

Comparison of group data to a reference group can be made without direct access to a data bank solely based on published

means and standard deviations of the reference data (e.g. Table 1 of this paper). All calculations are based on standard normal distribution data (z-values), which can be easily computed by subtracting the mean of the reference data and then dividing by the standard deviation of the reference data. *Z*-transformed reference data then have zero mean and unit standard deviation.

Comparison between records of test group data (= exp.) and a matched control group created as a fictitious subpopulation of reference group data of equal number (= con) is performed by *t*-test statistic (formula below). The distribution of *Z*-values of the control group is always given as mean = 0 and standard deviation (SD) = 1. (please note: equal group sizes are standard in case-control studies; see e.g. Gail 1998).

T-statistic for comparison:

$$t = (mean_{exp} - mean_{con})/squareroot(SD^2_{exp}/n_{cxp} + SD^2_{con}/n_{con}) \quad and: \quad mean_{con} = 0 \quad and \quad SD_{con} = 1 \quad and \quad n_{Exp} = n_{Con}$$

Appendix B. "Recipe" for practical conduct of statistical comparison

1. Calculation of tests is performed using simple probability calculators for t-tests (using e.g. STATISTICA Basic Statistics–Probability Calculator). Only mean, SD and number of data in the test group is needed. If respective statistical software is not available internet-based statistical freeware can be used (e.g. Quantitative Skills - SISA, see below)
2. All single data of the test group have to be transformed into a standard value (z-value) using mean and standard deviation (SD) from the appropriate age and gender cohort of the healthy subjects data base of the DFNS according to the following equation:

$$z = (\text{single subject}_{\text{exp. group}} - \text{mean}_{\text{control from reference data}})/\text{SD}_{\text{control from reference data}}$$

3. Input of mean and SD of test group data (exp.) and number of cases (e.g. $n = 32$)
4. Input of mean and SD of reference group data (con) and number of cases (e.g. $n = 32$) (always mean = 0 and SD = 1) and an equal number of cases (i.e. in this case, also $n = 32$)
5. Calculation of *t*-test (two sided and independent samples!!)
   a. you get a *t*-statistic with $2 \times (n - 1)$ degrees of freedom

Calculation by simple internet-based statistical software

Simple Interactive Statistical Analysis (SISA)

URL: http://www.quantitativeskills.com/sisa/ (as accessed 2009, July 16)

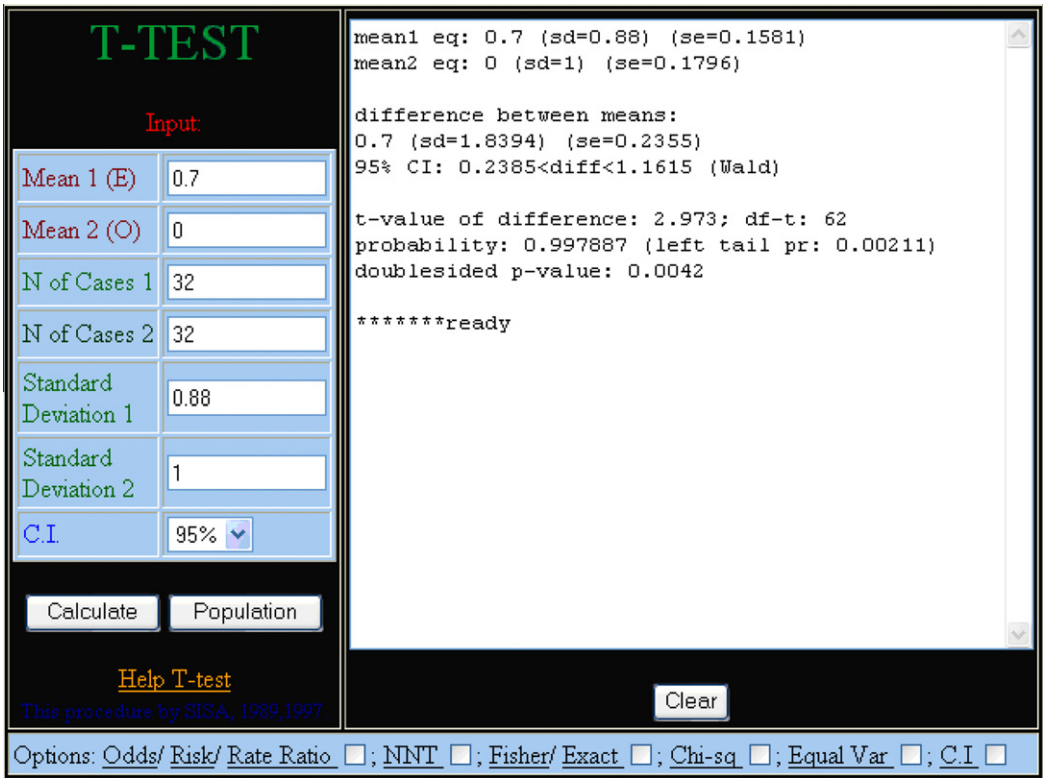
(general URL, there are many other statistical applications)

special application unpaired (whole sample) t-test

URL: http://www.quantitativeskills.com/sisa/statistics/t-test.htm

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.pain.2010.07.026.



References

[1] Blankenburg M, Boekens H, Hechler T, Maier C, Krumova E, Scherens A, Magerl W, Aksu F, Zernikow B. Reference values for quantitative sensory testing in children and adolescents: developmental and gender differences of somatosensory perception. Pain 2010;149:76–88.

[2] Box GEP, Jenkins GM. Time series analysis: forecasting and control. Revised ed. San Francisco: Holden-Day; 1997.

[3] Chesterton LS, Barlas P, Foster NE, Baxter GD, Wright CC. Gender differences in pressure pain threshold in healthy humans. Pain 2003;101:259–66.

[4] Cole LJ, Farrell MJ, Gibson SJ, Egan GF. Age-related differences in pain sensitivity and regional brain activity evoked by noxious pressure. Neurobiol Aging 2010;31:494–503.

[5] Cruccu G, Truini A. Tools for assessing neuropathic pain. PLoS Med 2009;6:e1000045.

[6] Edwards RR, Fillingim RB. Age-associated differences in responses to noxious stimuli. J Gerontol A Biol Sci Med Sci 2001;56:M180–5.

[7] Farrell M, Gibson S. Age interacts with stimulus frequency in the temporal summation of pain. Pain Med 2007;8:514–20.

[8] Farwell LA, Martinerie JM, Bashore TR, Rapp PE. Goddard PH optimal digital filters for long-latency components of the event-related brain potential. Psychophysiology 1993;30:306–15.

[9] Feine JS, Bushnell MC, Miron D, Duncan GH. Sex differences in the perception of noxious heat stimuli. Pain 1991;44:255–62.

[10] Fillingim RB, King CD, Ribeiro-DaSilva MC, Rahim-Williams B, Riley JL. Sex, gender, and pain: a review of recent clinical and experimental findings. J Pain 2009;10:447–85.

[11] Gail MH. Controls. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics, vol. 1. Chichester: Wiley; 1998. p. 932.

[12] Geber C, Scherens A, Pfau D, Nestler N, Zenz M, Tölle T, Baron R, Treede RD, Maier C. Procedure for certification of QST laboratories. Schmerz 2009;23:65–9 [Article in German].

[13] Gibson SJ, Farrell M. A review of age differences in the neurophysiology of nociception and the perceptual experience of pain. Clin J Pain 2004;20:227–39.

[14] Gibson SJ, Helme RD. Age-related differences in pain perception and report. Clin Geriatr Med 2001;17:433–56.

[15] Greenspan JD, Craft RM, LeResche L, Arendt-Nielsen L, Berkley KJ, Fillingim RB, Gold MS, Holdcroft A, Lautenbacher S, Mayer EA, Mogil JS, Murphy AZ, Traub RJ. Consensus working group of the sex, gender, and pain SIG of the IASP. Studying sex and gender differences in pain and analgesia: a consensus report. Pain 2007;132:S26–45.

[16] Hurley RW, Adams MC. Sex, gender, and pain: an overview of a complex field. Anesth Analg 2008;107:309–17.

[17] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale: Lawrence Erlbaum; 1988.

[18] Krauth J. Experimental design. A handbook and dictionary for medical and behavioural research. Amsterdam: Elsevier; 2000.

[19] Lachin JM. Sample size determination. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics, vol. 5. Chichester: Wiley; 1998. p. 3893–902.

[20] Lautenbacher S, Rollman GB. Sex differences in responsiveness to painful and non-painful stimuli are dependent upon the stimulation method. Pain 1993;53:255–64.

[21] Lin YH, Hsieh SC, Chao CC, Chang YC, Hsieh ST. Influence of aging on thermal and vibratory thresholds of quantitative sensory testing. J Peripher Nerv Syst 2005;10:269–81.

[22] Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance. A practical guide for biologists. Biol Rev Camb Phil Soc 2007;82:591–605.

[23] Petersen KL, Rowbotham MC. Quantitative sensory testing scaled up for multicenter clinical research networks: a promising start. Pain 2006;123:219–20.

[24] Pickering G, Jourdan D, Eschalier A, Dubray C. Impact of age, gender and cognitive functioning on pain perception. Gerontology 2002;48:112–8.

[25] Riley JL, Robinson ME, Wise EA, Myers CD, Fillingim RB. Sex differences in the perception of noxious experimental stimuli: a meta-analysis. Pain 1998;74:181–7.

[26] Rollman GB, Lautenbacher S. Sex differences in musculoskeletal pain. Clin J Pain 2001;17:20–4.

[27] Rolke R, Magerl W, Campbell KA, Schalber C, Caspari S, Birklein F, Treede RD. Quantitative sensory testing: a comprehensive protocol for clinical trials. Eur J Pain 2006;10:77–88.

[28] Rolke R, Baron R, Maier C, Tölle TR, Treede RD, Beyer A, Binder A, Birbaumer N, Birklein F, Bötefür IC, Braune S, Flor H, Huge V, Klug R, Landwehrmeyer GB, Magerl W, Maihöfner C, Rolko C, Schaub C, Scherens A, Sprenger T, Valet M, Wasserka B. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): standardized protocol and reference values. Pain 2006;123:231–43.

[29] Sarlani E, Greenspan JD. Gender differences in temporal summation of mechanically evoked pain. Pain 2002;97:163–9.

[30] Sarlani E, Grace EG, Reynolds MA, Greenspan JD. Sex differences in temporal summation of pain and after sensations following repetitive noxious mechanical stimulation. Pain 2004;109:115–23.

[31] Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study-review and new results. Stat Methods Med Res 2004;13:251–71.

[32] Torgén M, Swerup C. Individual factors and physical work load in relation to sensory thresholds in a middle-aged general population sample. Eur J Appl Physiol 2002;86:418–27.

[33] Treede RD, Baron R. How to detect a sensory abnormality. Eur J Pain 2008;12:395–6.

[34] Uitenbroek, DG. "SISA – Simple Interactive Statistical Analysis", 1997. Available: http://www.quantitativeskills.com/sisa/ [accessed 06.10.09]. Specific test available at: http://www.quantitativeskills.com/sisa/statistics/t-test.htm.

[35] Wang X, Zeng D, Seale H, Li S, Cheng H, Luan R, He X, Pang X, Dou X, Wang Q. Comparing early outbreak detection algorithms based on their optimized parameter values. J Biomed Inform 2010;43:97–103.

[36] Wilkinson L. APA task force on statistical inference. Statistical methods in psychology journals: guidelines and explanations. Am Psychol 1999;54:594–604.

[37] Zheng Z, Gibson SJ, Khalil Z, Helme RD, McMeeken JM. Age-related differences in the time course of capsaicin-induced hyperalgesia. Pain 2000;85:51–8.

[38] Zheng Z, Gibson SJ, Helme RD, McMeeken JM. The effect of local anaesthetic on age-related capsaicin-induced mechanical hyperalgesia – a randomised, controlled study. Pain 2009;144:101–9.

[39] Zumsteg D, Wieser HG. Effects of aging and sex on middle-latency somatosensory evoked potentials: normative data. Clin Neurophysiol 2002;113:681–5.