© 2009 The Authors
Journal compilation © 2009 Nordic Pharmacological Society. *Basic & Clinical Pharmacology & Toxicology*, **105**, 105–112

Doi: 10.1111/j.1742-7843.2009.00408.x

Applying Concepts of Generalizability Theory on Data from Experimental Pain Studies to Investigate Reliability

Assam Pryseley[1], Edouard Y Ledent[2], Asbjørn M Drewes[3,4], Camilla Staahl[3,4], Anne E Olesen[3,4] and Lars Arendt-Nielsen[4]

[1]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Limburg, Belgium, and Leuven Catholic University, Belgium, [2]GSK Biologicals, Rue de l'Institut, Rixensart, Belgium, [3]Department of Gastroenterology, Aalborg Hospital, Denmark, and [4]Department of Health Science and Technology, Aalborg University, Denmark

(Received 3 October 2008; Accepted 16 January 2009)

Abstract: This work demonstrates how full modelling power in statistically mixed models can be used to study generalizability (reliability) coefficients of advanced data from human experimental pain studies utilizing placebo data from drug screening trials. This can be used to help optimizing outcome parameters from existing data sets.

This study assesses the reliability of an experimental pain parameter based on electrical stimulation (single and repeated) of skin and muscle tissues and on chemical stimulation (6% saline) of muscle tissue. The data compiled were based on placebo data from a three × three cross-over designed study with repeated measurements within each period. The reliability coefficients of interest were: reliability between two measurements taken in an hour's duration within the same period (2SP) and reliability between two measurements taken at the same time points in different periods (2DP). The overall variation of the pain recording was characterized using the coefficient of variation (CV) and sample sizes were estimated for comparison purposes.

Generally, the pain measurements showed a moderate overall variability with a mean coefficient of variation value of 32% (range 14–39%) with the lowest variation for the temporal summation data (14%). This renders these tests useful for analgesic testing recruiting useful sample sizes in cross-over design (N = 8 for temporal summation from skin, N = 22 for temporal summation from muscles, 95%), but may be unhandy in parallel design (N = 14 and N = 48, respectively, 95%). For both single electrical stimulation and chemical stimulation, a higher number of volunteers are needed in both cross-over and parallel designs.

Pain is a multidimensional unpleasant sensory and emotional experience and cannot as such be represented or described by a single parameter or number. Hence, assessing pain quantitatively is a challenging task. However, different possibilities in human experimental pain research exist to assess quantitatively various aspects of and mechanisms involved in the complex sensory experience of pain.

Human experimental pain research involves two separate topics: (i) Standardized activation of the nociceptive system; and (ii) Assessment of the evoked responses. The ultimate goal of modern pain assessment procedures is to obtain a better understanding of mechanisms involved in pain transduction, transmission, and perception under normal and pathophysiological conditions. Such a mechanism-based approach can provide better characterization, prevention and management of pain. In recent years, human experimental pain models have been developed as bio-markers to be used in early drug development for screening of analgesic potency of new and existing analgesics [1].

In experimental studies designed for screening analgesics, human experimental pain models provide a means to overcome some factors that confound clinical drug trials. These models allow the investigator to control the nature, localization, intensity, frequency and duration of the induced pain stimulus, and provide standardized quantitative measures of a psychophysical response [2]. However, most experimental pain stimuli cannot mimic clinical pain but may act as a surrogate model and hence act as a biomarker for analgesic efficacy in patients. The use of multi-modal, multi-tissue differentiated experimental pain stimulation provides a possibility to differentiate the pain responses and provide a mechanism-based evaluation [3] of analgesic efficacy.

In drug screening trials, it is mandatory that the pain tests are reproducible, valid and respond in a uniform way to changes in nociceptive excitability [1]. Many clinical pain measures are based on clinicians' or patients' subjective observations, whereas the experimental measures theoretically are more stable because the nociceptive input (stimulus intensity) can be standardized. Clinical as well as experimental pain measures are, however, prone to errors and variability, some owing to their subjective nature. Before such measures are ready for application in drug trials, they should be highly reliable to avoid type I and type II errors.

Author for correspondence: Lars Arendt-Nielsen, Department of Health Science and Technology, Center for Sensory-Motor Interaction, Aalborg University, Fredrik Bajers Vej 7, 9220 Aalborg E, Denmark (fax + 45 98154008, e-mail lan@hst.aau.dk).

Reliability coefficients express the ability to differentiate among subjects. They are ratios of variances: in classical terms, the variance attributed to the difference among subjects, divided by the total variance [4]. Such parameters should be known and minimized, and practical as well as theoretical considerations should be optimized to be able to perform the most reliable clinical trials.

As stated by Fleiss [5]: 'The most elegant design of a clinical study will not overcome the damage by unreliable or imprecise measurement.' In clinical trials, one typically wants to differentiate among treatments. If reliability is low, the ability to differentiate between the different subjects in the different treatment arms decreases. One of the consequences of unreliability described by Fleiss is an increase in sample size for trials with a primary parameter exhibiting low reliability.

The classical theory behind the estimation of reliability can be extended to the Generalizability Theory (GT) by estimating the magnitude of multiple sources of measurement error and providing reliability and generalizability coefficients tailored to the proposed use of the measurement and isolating major sources of error so that a cost-efficient measurement design can be built [6]. By investigating other sources of error, the clinical trialist could learn about performance of scales or other measurements in certain subgroups and the impact of such factors on reliability.

This article aims at applying the concepts of the GT on selected experimental pain measures used in drug screening trials (data from the study of Olesen *et al.* [7]) with focus on the sources of variance and their impact on the reliability and generalizability of the measurements. This information is important for powering cross-over and parallel experimental drug screening pain trials.

Materials and Methods

Study subjects. Eighteen healthy, non-smoking, volunteers aged between 18 and 30 years with a body mass index within the range of 18–30 kg/m^2 completed the study. All subjects were in good general health with no clinically relevant abnormalities of medical history, physical examination, clinical or laboratory evaluation. The subjects were informed about the risk of the study and were paid for participating. Oral and written informed consent was obtained from all subjects. The Ethics Committee of Northern Jutland (VN 2004/62) and the Danish Medicines Agency (EUDRACT nr. 2004-002,605-77) approved the study [7].

Study design. This was a single-centre, double-blind, randomised, three-way cross-over study in healthy subjects. The experimental tests included single and repeated electrical stimulation of the skin and muscle, and intra-muscular saline-evoked muscle pain. Pain intensity assessments were carried out using a visual analogue scale (VAS), pre and post dose. Subjects were screened and underwent familiarization of the experimental tests within 14 days of the initial treatment visit. They then returned for three treatment visits, each separated by a minimum of 3 days. Experimental testing commenced 30 min. before dosing (baseline) and 1, 2, and 3 hrs after dosing. Muscle stimulation with saline was only done at baseline and 2 hrs after dosing.

Study treatments. Randomized subjects were assigned to receive one of the three treatments during each period of this cross-over study with the order of treatment dosing randomized by the use of a Latin square design. The treatments were: (i) Drug A, (ii) Drug B or (iii) Placebo. All treatments were supplied unmarked and blister-packed. Medications were taken orally with 150 ml of water with subjects fasting for 3 hrs before dosing. For the present article, all data were used for further analysis in order to optimise benefits from GT.

Pain stimulations

During the first visit, each subject went through all pain tests listed below to ensure that they understood and were familiarized with the rating procedures and the nature of the stimuli. They were instructed how to use the continuous VAS scale where the minimum was 0 = no perception and the maximum 10 = unbearable pain intensity. The VAS is simple and efficient to use and has been considered by many to be reliable as a ratio-scale measure of pain intensity [8].

Single electrical stimulation.

Skin: A computer-controlled constant current stimulator (University of Aalborg, Denmark) delivered a 25 msec., train-of-five, 1 msec. square-wave impulse (perceived as a single stimulus) to the skin over the sural nerve using two bipolar Ag/AgCl-electrodes (Neuroline, Medicotest A/S, Ølstykke, Denmark). The surface of the electrodes was 15×10 mm, and the distance between the two electrodes was 15 mm. The current intensity was increased from 1 mA in increments of 0.5 mA until pain detection threshold. The pain detection threshold is defined as the intensity when the perceived sensation changes from a mechanical sensation to pain.

Muscle: Similar current pulses were delivered to the anterior tibialis muscle by needle electrodes (30×0.35 mm; Dantec, Denmark) fully inserted into the muscle belly, placed 1 cm perpendicularly apart to the surface, starting with a stimulus intensity of 0.5 mA for the single stimulus. The current intensity was increased stepwise with 0.25 mA until pain was evoked (pain detection threshold).

Repeated electrical stimulation (temporal summation).

Skin: A 25 msec., train-of-five, 1 msec. square-wave constant current pulses was used. This stimulus burst was repeated five times with a frequency of 2 Hz. The current intensity was increased from 1 mA in steps of 0.5 mA, until a subjective pain sensation was evoked (pain detection threshold) during the train of stimuli.

Muscle: The procedures described for the skin was repeated for the muscle electrodes, starting with a stimulus intensity of 0.5 mA.

Stimulation of the muscle with hypertonic saline.

Muscle pain was induced by injection of hypertonic saline (5% NaCl). A computer-controlled syringe pump (IVAC, model 770, USA) was connected through an extension tube (IVAC, G30303, extension set with polyethylene inner line) to a stainless disposable needle (27G, 40 mm). The needle was introduced in the left side anterior tibialis muscle, 14 cm distally from the caudal end of the patella, 2 cm laterally to the anterior edge of the tibia, and 20 mm in depth. A measure of 0.5 ml of 5% hypertonic saline was administered over 20 sec. Subjects rated pain intensity continuously for 7 min. on the electronic VAS. The area below the VAS curve (AUC-VAS) was calculated as well as the time to peak pain (TP-VAS) and the maximum VAS score (MAX-VAS).

Statistics

To eliminate errors relating to differences between periods in baseline pain recordings, the change in stimulus intensity relative to baseline was used in the calculations.

Reliability and generalizability.

In this section, the GT of Cronbach *et al.* [9] to estimate various reliability coefficients of interest is introduced. The following subsection uses the method of Vangeneugden *et al.* [10], who illustrated how full modelling power in mixed models can be used to study generalizability.

In classical test theory, the outcome of a test is modelled as two random variables

$$Y = \tau + \varepsilon, \tag{1}$$

where Y represents an observation or measurement, τ is the true score, ε the corresponding measurement error and $\mathrm{Var}(Y) = \mathrm{Var}(\tau) + \mathrm{Var}(\varepsilon)$. It is assumed that the measurement errors are mutually uncorrelated as well as with the true score. The reliability (R) of a measuring instrument is defined as the ratio of the true score variance to the observed score variance, that is,

$$R = \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(\tau) + \mathrm{Var}(\varepsilon)} \tag{2}$$

In the case of two parallel measurements, we have $Y_1 = \tau + \varepsilon_1$ and $Y_2 = \tau + \varepsilon_2$, with $\mathrm{Var}(Y_1) = \mathrm{Var}(Y_1) = \mathrm{Var}(Y)$ and $\mathrm{Var}(\varepsilon_1) = \mathrm{Var}(\varepsilon_2) = \mathrm{Var}(\varepsilon)$. Therefore, the reliability of the two measurements equals

$$\begin{aligned} R = \mathrm{Corr}(Y_1, Y_2) &= \frac{\mathrm{Cov}(Y_1, Y_2)}{\sqrt{\mathrm{Var}(Y_1)}\sqrt{\mathrm{Var}(Y_2)}} \\ &= \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(\tau) + \mathrm{Var}(\varepsilon)}. \end{aligned} \tag{3}$$

Classical test theory assumes that an observation is a combination of an individual's true score and random measurement error. The assumption that all variance in scores can be divided into true and error variance is rather simplistic. The essence of the GT is the recognition that in any measurement situation, there are multiple sources of error variance. The goal is to attempt identification, measurement and thereby possibly to find strategies to reduce the influence of these sources on the measurement in question. Thus, it is more efficient to investigate all the sources of variability in a single study using all the data to estimate the variance between the subjects and the various components of error variance. This can provide a lot of information on observer reliability and can determine the relative importance of each component. The reliability estimates in this report are based on absolute decisions [10]. This ensures that the total variance from the model is used to calculate the reliability measures.

By reasonably identifying the most likely sources of error in a measurement, we have defined our 'universe' of generalization. Generalizability coefficients will depend on which universe, which variance components are considered and which factors are allowed to vary and which remain fixed. If the sources that we have identified are trivial, and we have missed some important sources of error, then there will be a large amount of variance due to random error. This may lead to reliability estimates which are biased downwards, that is, provide low estimates for the reliability coefficients.

Reliability and generalizability via linear mixed models.

Based on the repeated measurements taken on the subjects within periods of the same study, possible between-subject and period variability and, the hierarchical nature of the data, a linear mixed-effects framework is considered. A linear mixed model is any model which satisfies [11–13]:

$$\begin{aligned} &Y_i = X_i\beta + Z_i b_i + W_i + \varepsilon_i \\ &b_i \sim N(0, D), \\ &W_i \sim N(0, \tau^2 H_i) \\ &\varepsilon_i \sim N(0, \Sigma_i), \\ &b_1, \ldots, b_n, \varepsilon_1, \ldots, \varepsilon_n \text{ independent,} \end{aligned} \tag{4}$$

where Y_i is the n_i-dimensional response vector for subject i, $1 \le i \le N$, N is the number of subjects, X_i and z_i are $(n_i \times p)$- and $(n_i \times q)$- dimensional matrices of known covariates, β is a p-dimensional vector containing the fixed effects, b_i is the q-dimensional vector containing the random effects, and ε_i is an n_i-dimensional vector of residual components. D is a general $(q \times q)$ covariance matrix with (i, j) element $d_{ij} = d_{ji}$ and $\sum_i$ is a $(n_i \times n_i)$ covariance matrix which depends on i only through its dimension n_i, that is, the set of unknown parameters in $\sum_i$ will not depend upon i. Serial correlation is captured by the realisation of a Gaussian stochastic process, W_i.

Model (4) is very general and flexible given that it permits one to estimate different variances; allowing for instance a different variance for the measurement at each time point. Additionally, fixed effects can be accommodated. This flexibility enables the calculation of reliability and generalizability coefficients from data resulting from clinical trials. Based on equations (2) and (3), a general formula to calculate reliability can be derived from the linear mixed model (4). Denote by Y_{jt} the observed measurement of subject i at time point t. Denote the test-retest reliability between time points s and t by $R\,(s,t)$, we have

$$\begin{aligned} R &= \mathrm{Corr}(Y_{is}, Y_{it}) \\ &= \frac{z_s D z_t' + \tau^2 (H_i)_{st}}{\sqrt{z_s D z_s' + \tau^2 + \sigma^2}\,\sqrt{z_t D z_t' + \tau^2 + \sigma^2}}. \end{aligned} \tag{5}$$

The reliability coefficient is mostly derived via the intraclass correlation. As Bartko [14] has demonstrated, the intraclass correlation calculated as a ratio of variances that are estimated by means of a linear model is only correct when it can be interpreted as a correlation coefficient. It can be shown that equation (5) can be derived as a conditional correlation coefficient [10]. Thus, equation (5) can be used to derive reliability and various generalizability coefficients for the linear mixed model framework. Intraclass correlation >0.80 is generally considered excellent [15], but intraclass correlation >0.60 is acceptable for many purposes in experimental pain studies [7].

The coefficient of variation reflects the overall variability of a model and was used to characterize the overall variance. The coefficient of variation was used to estimate sample sizes needed for parallel and cross-over studies for comparison purposes. The desired power of the analysis was set to 0.95 and α to 0.05. The estimation was done for the detection of an analgesic effect producing a 30% decrease or increase of the pain recording, which has been shown to be realistic [16]. Sample sizes below 30 are manageable in advanced experimental pain models. Larger sample sizes may make the study impractical and expensive to perform [16].

Let Y_{ipk} denote the kth, $1 \leq k \leq K$ measurement in period p, $1 \leq p \leq P$, of a response on subject i, $1 \leq i \leq N$. Also, let B_{ip} be a baseline measurement in period p for subject i. A particular realisation of model (4) which will be fitted to our data sets is given by:

$$D_{ipk} = \text{mean}_{ipk} + b_{0i} + \ b_{1p} + \varepsilon_{ipk}, \qquad (6)$$

where $D_{ipk} = Y_{ipk} - B_{ip}$ and mean_{ipk} represent the fixed effects from the grand mean, baseline, period, treatment, time, baseline × time and treatment × time. b_{0i} and b_{1p} represent the between subject variability and between period variability with $\text{Var}(b_{0i}) = d_0$ and $\text{Var}(b_{1p}) = d_1$, respectively. We assume that the errors, $\varepsilon_{1pk}, \ldots, \varepsilon_{npk}$, are independent across periods but correlated within periods. Errors within a period follow an autoregressive process modelled using a spatial power function such as $\sum_i = \sigma_e^2 \rho^{d(k,s)}$ where σ_e^2 is the residual variance [10,11]. The correlation between two repeated measurements on a subject within a period is given by $\rho^{d(k,s)}$. $d(k,s)$ is the Euclidean distance between time points with $d(k,s) = 0$ and $d(k,s) > 0$.

Useful reliability coefficients of interest are the reliability between measurements taken on the same time point across different periods (2DP) and the reliability between two measurements taken in an hour's duration within the same period (2SP). Based on model (6), these reliability coefficients and the coefficient of variation (CV*)* can be calculated as follows:

$$2\text{DP} = R(Y_{ipk}, Y_{ip'k}) = \frac{d_0}{d_0 + d_1 + \sigma_e^2};$$

$$2\text{SP} = R(Y_{ipk}, Y_{ipk'}) = \frac{d_0 + d_1 + |\rho^{d(k,k')}|\sigma_e^2}{d_0 + d_1 + \sigma_e^2}, \text{ and}$$

$$CV = \frac{\sqrt{d_0 + d_1 + \sigma_e^2}}{\mu}.$$

SAS version 9.1.3, specifically The Mixed Procedure, was used for the estimation of reliability and generalizability coefficients. The delta method can be used to obtain standard errors and hence confidence intervals for the reliability estimate. The variances based on the delta method are given as:

$$\text{Var}(2\text{DP}) = \frac{(d_1 + \sigma_e^2)^2\text{Var}(d_0) + d_0^2(\text{Var}(d_1) + \text{Var}(\sigma_e^2))}{(d_0 + d_1 + \sigma_e^2)^4}$$

$$\text{Var}(2\text{SP}) = \frac{\begin{array}{l}[\sigma_e^2(1 - |\rho|)]^2(\text{Var}(d_0) + \text{Var}(d_1)) \\ + [\sigma_e^2(d_0 + d_1 + \sigma_e^2)]^2\text{Var}(\rho) \\ + [(d_0 + d_1)(|\rho| - 1)]^2\text{Var}(\sigma_e^2)\end{array}}{(d_0 + d_1 + \sigma_e^2)^4}.$$

NQUERY and PASS2005 were used for sample size calculations based on 2 × 2 cross-over and parallel designs. Calculating sample size for a parallel design is trivial; thus, we focus on sample size calculation for cross-over design. This can be done using standard sample size package for t-test procedure using MSE/2 as the estimated variance, to obtain the number of subjects per sequence. Using the cross-over package in PASS2005 and the standard deviation as mentioned in table 1 yields required sample sizes for a 2 × 2 cross-over design.

Results

In this section, we present the result obtained from applying the linear mixed model (formula 6) on the motivational data as well as estimates of generalizability coefficients and sample sizes. About 0–9.2% of the observations (assessments) were missing for some of the pain recordings due to technical problems. The percentage of missingness is <10% and the analyses based on linear mixed models are valid under the so-called missing at random assumption [12,13,17], which means, given the observed data and measured covariates, there is no further information in the missing data regarding the process that governs missingness. Thus, neither further techniques to account for missingness were, nor have to be, employed. Also, some observations were excluded from the analyses because they were identified as influential outliers; depending on pain recordings the number of outlying observations ranged between 0 and 6 (4%), mostly assessments on different patients from different periods.

For outcomes derived from response to chemical muscle stimulation, that is, AUC-VAS, MAX-VAS, and TP-VAS, models including between period variability could not be brought to convergence. Furthermore, for the other outcomes based on electrical stimulation, likelihood ratio tests derived from mixtures of Chi-squares [12,13] indicated lack of evidence for substantial between period variability. Even so, reliability estimates calculated for the later outcomes with and without the between subject variability (i.e. $d_1 = 0$) indicate that the periods have very little impact on the reliability of the measurements (table 2). Thus, henceforth, we present results from models without the between period

Table 1.
Sample sizes required to detect an analgesic effect of producing a 30% decrease of the average pain recordings with a significance level of 0.05 and power set to both 0.90 and 0.95, in a cross-over and a parallel design are presented. These sample sizes are for comparison purposes as a 30% reduction in the specific endpoint may not be of clinical relevance. The standard deviations and magnitude of a 30% reduction of the mean value for each pain recording are also presented. Change to be detected holds the corresponding 30% reduction of the overall mean of a given Response. Standard deviation represents the corresponding standard deviation obtained from the model required for testing treatment effect. Cross-over and Parallel show the required number of subjects for cross-over and parallel designs, respectively, based on 90% and 95% power. PDT = Pain detection threshold.

Tissue	Response, stimulus	Change to be detected	Within standard deviation	Total standard deviation	Cross-over		Parallel	
					90%	95%	90%	95%
Skin	PDT, Single electrical (mA)	12.03	11.30645	11.93063	22	26	44	54
	PDT, Temporal electrical (mA)	1.66	0.66296	0.78740	8	8	12	14
Muscle	PDT, Single electrical (mA)	11.24	13.97856	14.39583	36	44	72	88
	PDT, Temporal electrical (mA)	0.75	0.646265	0.69282	18	22	38	48
	AUC-VAS (arbitrary units)	163.45	170.9703	202.21450	26	32	68	82
	MAX-VAS (arbitrary units)	11.80	12.79742	15.01832	28	34	72	88
	TP-VAS (sec)	26.79	31.47948	33.95394	32	38	70	86

variability. The generalizability coefficients and coefficients of variations for each pain recording are listed in table 3. Table 1 presents the calculated sample sizes required to detect a 30% change of pain parameter where the value of power is set to 90% and 95%.

Single electrical stimulation.

Skin – pain detection threshold. For single electrical stimulation on the skin, the pain threshold showed a relatively high reproducibility within periods and a low reproducibility across periods. This can be seen from the high reliability between two measurements taken in an hour's duration within the same period (2SP = 0.78) and the low reliability between two measurements taken at the same time point in different periods (2DP = 0.10) (table 3). This showed that the volunteers were able to reproduce their own pain recordings within periods but could not do so across periods. The pain threshold had a moderate coefficient of variation value (31%), which renders it useful in the testing of analgesics for cross-over experimental design but require moderate sample sizes for parallel experimental design (table 1).

Muscle – pain detection threshold. When single electrical stimulation was applied to the muscle, the pain detection threshold showed high reproducibility within the same periods (2SP = 0.75) and very low reproducibility across different periods (2DP = 0.07). It also showed a moderate overall variability with a coefficient of variation value of 39% (table 3). Moderate sample sizes are required for a cross-over study unlike for a parallel study which requires a large number of patients rendering such studies unhandy (table 1). This test may be useful for testing analgesic effects for cross-over studies.

Repeated electrical stimulation (temporal summation).

Skin – summation pain detection threshold. Pain detection threshold for temporal electrical stimulation on the skin showed moderate and low reproducibility within periods (2SP = 0.54) and between periods (2DP = 0.21), respectively. However, the coefficient of variation value is low (14%) indicating a small overall variation (table 3). Furthermore, very realistic sample sizes are required for both cross-over (N = 8, 95%) and parallel (N = 14, 95%) studies (table 1). Thus, the test will be sensitive to analgesic modulation for both cross-over and parallel experimental designs.

Table 2.
Reliability (2DP [two measurements at different periods] and 2SP [two measurements at same period]) estimates when between period variability is considered and ignored in the modelling process. Period Effect assumes that there is extra variability induced by the different periods in the cross-over design, implying that the parameter d_1 is to be estimated in the modelling process. No Period Effect assumes that there is no extra variability induced by the different periods in the cross-over design and $d_1 = 0$ in the modelling process. PDT = Pain detection threshold.

		2DP		2SP	
Tissue	Response, stimulus	With between period variability	Without between period variability	With between period variability	Without between period variability
Skin	PDT, Single electrical (mA)	0.096	0.098	0.769	0.784
	PDT, Temporal electrical (mA)	0.256	0.210	0.335	0.541
Muscle	PDT, Single electrical (mA)	0.065	0.072	0.742	0.748
	PDT, Temporal electrical (mA)	0.137	0.152	0.642	0.649

Table 3.

Results obtained from the statistical analysis of the various pain recordings from skin and muscle. Subject represents the between subject variability. Corr holds the value for the correlation between successive repeated measurements within a period of the study. Residual shows the residual variance for each pain recording. 2DP (two measurements at different periods) and 2SP (two periods at same period) represent the reliability between two measurements taken at the same time point in different periods and the reliability of two measurements taken an hour apart within the same period, respectively. Coefficient of variation is the estimate for the coefficient of variation based on the total variance of each pain recording. PDT = Pain detection threshold.

Tissue	Response, stimulus	Subject	Corr	Residual	2DP	2SP	CV (%)
Skin	PDT, Single electrical (mA)	15.55	0.76	143.04	0.10	0.78	31
	PDT, Temporal electrical (mA)	0.14	0.42	0.51	0.21	0.54	14
Muscle	PDT, Single electrical (mA)	15.64	0.73	200.88	0.07	0.75	39
	PDT, Temporal electrical (mA)	0.09	0.59	0.51	0.15	0.65	30
	AUC-VAS (arbitrary units)	11659.86	−0.15	29230.84	0.29	0.39	37
	MAX-VAS (arbitrary units)	61.78	0.15	163.77	0.27	0.39	38
	TP-VAS (sec)	161.91	0.22	990.96	0.14	0.33	38

Muscle – summation pain detection threshold. When the muscle was stimulated using temporal electrical stimulation, the pain detection threshold showed a high reproducibility within the same periods (2SP = 0.65) and a low reproducibility across different periods (2DP = 0.15). It also showed a moderate overall variability with a coefficient of variation value of 30% (table 3). The sample sizes required to detect a 30% reduction in the pain detection threshold for cross-over studies are reasonable (N = 22, 95%), but less useful sample sizes are required for parallel studies (N = 48, 95%), (table 1). The high within period reproducibility may render this test useful for testing analgesic effects in cross-over studies.

Intramuscular hypertonic saline.

AUC-VAS. The area under the VAS curve was not reproducible, neither within nor between periods (table 3). The overall variability was 37%. As a consequence, the required sample size for detecting a 30% modulation of the pain response in a cross-over study was 32 and 82 for parallel studies (table 1).

MAX-VAS. The maximum VAS score was not reproducible, neither within nor between periods (table 3). A coefficient of variation value of 38% indicated a moderate overall variation. Hence, the sample sizes required to detect a 30% modulation in a cross-over study were 34 and 88 for parallel studies (table 1).

TP-VAS. The time to peak VAS was not reproducible, neither within nor between periods (table 3). A coefficient of variation of 38% indicated a moderate overall variation. Hence, the sample sizes required to detect a 30% modulation in the response for the cross-over study were 38 and 86 for parallel studies (table 1).

Discussion

The present study showed that particular temporal summation assessed from the skin is a highly reliable parameter that can be used in both cross-over and parallel studies with a sample size of 8 and 14, respectively (able to detect a 30% decrease, P < 0.05, power 0.95). Temporal summation is a very important and potent pain mechanism. In humans, central integration manifested as temporal summation (increased pain reaction in volunteers or patients to a train [e.g. 5 pulses] of repeated [e.g. 2 Hz] stimuli) is assumed to represent the initial phase of the wind-up process as seen in animals [18–21] and represents relevant and potent mechanisms in central sensitization.

Temporal summation is facilitated in experimentally induced hyperalgesic areas [22] and in patients with different chronic pain conditions (neuropathic, musculoskeletal) [23–25]. Temporal summation is difficult to block pharmacologically but drugs efficient in inhibiting central sensitization (in e.g. neuropathic pain) also block temporal summation [26,27].

Multi-modal, multi-tissue-differentiated experimental pain models offer a unique opportunity of comprehensive assessment of effects of analgesics and reveal pharmacological insight into new and existing compounds. This may help the development and targeting of new analgesics. The experimental pain model can be used in either cross-over or parallel studies, and understanding the variability for different experimental pain parameters is therefore important.

The assessment of observer reliability is essential for the interpretation of medical observations both in fundamental research as well as in medical practice [28]. Cronbach *et al.* [9] devised the GT, which essentially recognises multiple sources of error variance in any measurement situation and hence can be used to estimate multiple reliability coefficients. Shavelson *et al.* [6] and Dunn [29] made a plea for more extensive use of GT studies. Shavelson *et al.* [6] surmised that one of the reasons why GT studies are not widely used is the cost to set up generalizability studies.

Following the ideas of Vandeneugden *et al.* [10], we used clinical trial data (placebo) to estimate (reliability) generalizability coefficients using linear mixed models. This approach measures the contribution of different sources of variation to the total measurement error, after accommodating fixed effects, the most prominent one being the treatment effect.

The goal is to identify important sources of variability in a measurement situation from the outset and attempt to quantify these sources of error. Our motivational data were based on a three × three cross-over design with repeated measurements within each period. We assumed the following sources of variance: between period variability, between subject variability, serial correlation between the measurement within a period for a given subject, and within subject variability (residual variance). Based on these sources of variability, the following reliability coefficients were of interest: reliability between two measurements taken in an hour's duration within the same period (2SP) and the reliability between two measurements taken at the same time point in different periods (2DP).

The within-subject variation makes up a dominant part of the total variability for all assessment parameters. We should allow for the fact that some sources of variability might have gone unnoticed. One also has to consider the fact that including too many sources of error makes the mixed models complicated, and getting such models to converge is not always a trivial task. Hence, it is important to identify important and meaningful sources of error to ensure convergence of the mixed models and obtain stable estimates of the variance parameters.

As expected, the reliability between two measurements taken an hour apart within a period is consistently considerably higher than the reliability between two measurements taken at the same time point between different periods. This probably reflects the fact that the various devices and stimuli are difficult to place in exactly the same position between periods [30].

Conclusion

Although validation studies can be initiated, it may be more practicable to use existing clinical trial data for assessing and estimating reliability coefficients of interest, after correcting for fixed effects (the most prominent being treatment effect), using GT. The present results also indicate that the usual reliability studies, using only one period without administration of treatments, lack the ability to capture the very low reliability across periods.

In settings where the within subject variability is much less than the total variability, the 'gain' in using cross-over is much greater in terms of reduction in the number of subjects. However, the within subject variability constitutes a dominant part of the total variability for this study. Consequently, twice the number of subjects in a cross-over is not quite different as compared to a parallel design. In spite of this, the investigator may prefer to have longer duration trials as compared to enrolling a little more than twice the number of subjects for parallel studies.

We have illustrated that other sources of variability may have an impact on the reliability of an outcome and that reliability studies with one period miss to capture the low reliability across periods. Therefore, performing pilot cross-over trials is the appropriate way to assess reliabilities if the investigator considers using such experimental designs. Essentially, the administration of treatments during these pilot studies does not prevent the estimation of reliabilities in an appropriate and sensible way.

References

1 Arendt-Nielsen L, Curatolo M, Drewes A. Human experimental pain models in drug development: translational pain research. Cur Opin Inv Drugs 2007;**8**(1):41–53.

2 Arendt-Nielsen L. Induction and assessment of experimental pain from human skin, muscle and viscera. In: Jensen TS, Turner JA, Wiesenfeld-Hallin Z (eds.) Proceedings of the 8th World Congress on Pain, Vancouver, Canada, August 17–22: Progress in Pain Research and Management. IASP Press, Seattle. 1997;**8**:393–425.

3 Arendt-Nielsen L, Graven-Nielsen T. Central sensitisation in fibromyalgia and other musculoskeletal disorders. Curr Pain Headache Rep 2003;**7**:355–61.

4 Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing inter-rater reliability. Psychol Bull 1979;**86**:420–8.

5 Fleiss JL. Design and Analysis of Clinical Experiments. Wiley, New York 1986.

6 Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. Am Psychol 1989;**44**:922–32.

7 Olesen AE, Staahl C, Ali Z, Drewes AM, Arendt-Nielsen L. Effects of paracetamol combined with dextromethorphan in human experimental muscle and skin pain. Basic Clin Pharmacol Toxicol 2007;**101**(3):172–6.

8 Bijur PE, Silver W, Gallagher EJ. Reliability of the visual analog scale for measurement of acute pain. Acad Emerg Med 2001;**8**:1153–7.

9 Cronbach LJ, Rajaratnam N, Gleser GC. Theory of generalizability: A liberalization of reliability theory. Br J Stat Psychol 1963;**16**:137–63.

10 Vangeneugden T, Laenen A, Geys H, Renard D, Molenberghs G. Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. Biometrics 2005;**61**:295–304.

11 Laird NM, Ware JH. Random effects models for longitudinal data. Biometrics 1982;**38**:963–74.

12 Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. Springer-Verlag, New York 2000.

13 Molenberghs G, Verbeke G. Models for Repeated Discrete Data. Springer-Verlag, New York 2005.

14 Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966;**19**:3–11.

15 Grafton KV, Foster NE, Wright CC. Test-retest reliability of the Short-Form McGill Pain Questionnaire: assessment of intraclass correlation coefficients and limits of agreement in patients with osteroarthritis. Clin J Pain 2005;**21**:73–82.

16 Staahl C, Reddy H, Andersen SD, Arendt-Nielsen L, Drewes AM. Multi-modal and tissue-differentiated experimental pain assessment: reproducibility of a new concept for assessment of analgesics. Basic Clin Pharmacol Toxicol 2006;**98**:201–11.

17 Molenberghs G, Kenward MG. Missing Data in Clinical Studies. John Wiley & Sons, New York 2007.

18 Price DD, Hayes RL, Ruda M, Dubner R. Spatial and temporal transformations of input to spinothalamic tract neurons and their relation to somatic sensations. J Neurophysiol 1978;**41**(4):933–47.

19 Arendt-Nielsen L, Brennum J, Sindrup S, Bak P. Electrophysiological and psychophysical quantification of temporal summation

in the human nociceptive system. Eur J Appl Physiol 1994;**68**:266–73.

20 Arendt-Nielsen L, Sonnenborg FA, Andersen OK. Facilitation of the withdrawal reflex by repeated transcutaneous electrical stimulation: an experimental study on central integration in humans. Eur J Appl Physiol 2000;**81**(3):165–73.

21 You HJ, Dahl Morch C, Chen J, Arendt-Nielsen L. Simultaneous recordings of wind-up of paired spinal dorsal horn nociceptive neuron and nociceptive flexion reflex in rats. Brain Res 2003;**960**(1–2):235–45.

22 Arendt-Nielsen L, Andersen OK, Jensen TS. Brief, prolonged and repeated stimuli applied to hyperalgesic skin areas: a psychophysical study. Brain Res 1996;**712**(1):165–7.

23 Gottrup H, Kristensen AD, Bach FW, Jensen TS. Aftersensations in experimental and clinical hypersensitivity. Pain 2003;**103**(1–2):57–64.

24 Sørensen J, Graven-Nielsen T, Henriksson KG, Bengtsson M, Arendt-Nielsen L. Hyperexcitability in fibromyalgia. J Rheumatol 1998;**25**:152–5.

25 Curatolo M, Petersen-Felix S, Arendt-Nielsen L, Giani C, Zbinden AM, Radanov BP. Central hypersensitivity in chronic pain after whiplash injury. Clin J Pain 2001;**17**(4):306–15.

26 Arendt-Nielsen L, Petersen-Felix S, Fischer M, Bak P, Bjerring P, Zbinden AM. The effect of N-methyl-D-aspartate antagonist (ketamine) on single and repeated nociceptive stimuli: a placebo-controlled experimental human study. Anesth Analg 1995;**81**(1): 63–8.

27 Felsby S, Nielsen J, Arendt-Nielsen L, Jensen TS. NMDA receptor blockade in chronic neuropathic pain: a comparison of ketamine and magnesium chloride. Pain 1996;**64**(2):283–91.

28 Armitage P, Colton T. Encyclopedia of Biostatistics. Wiley, New York 1998.

29 Dunn G. Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors. Arnold London 1989.

30 Rosier EM, Iadarola MJ, Coghill RC. Reproducibility of pain measurement and pain perception. Pain 2002;**98**:205–16.