ABSTRACT: Conduct of a large, multicenter trial of the aldose reductase inhibitor zenarestat provided data on the reproducibility of multiple electrophysiologic (nerve conduction studies, NCS) and quantitative sensory (QST) tests. Baseline and 12-month electrophysiologic data from approximately 1100 patients at multiple centers were available for analysis. Intersite variability contributed minimally to overall test variance. All NCS tests were highly reproducible. Cool thermal and vibration QST thresholds, as measured by CASE IV instrumentation, were also highly reproducible. Intersubject variance accounted for the majority of variance for all parameters measured. Repeating NCS and QST measures decreased sample sizes needed to show statistical significance. Consideration of these observations, particularly with regard to QST, should aid in the design of future clinical trials investigating neuropathy.

Muscle Nerve **34:** 214–224, 2006

VALUE OF REPEATED MEASURES OF NERVE CONDUCTION AND QUANTITATIVE SENSORY TESTING IN A DIABETIC NEUROPATHY TRIAL

SHAWN J. BIRD, MD,[1] MARK J. BROWN, MD,[1] CATHIE SPINO, DSc,[2] SHARON WATLING, PharmD,[2] and HOWARD L. FOYT, MD, PhD[2]

[1] Department of Neurology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA
[2] Pfizer Global Research and Development, Ann Arbor Laboratories, 2800 Plymouth Road, Ann Arbor, Michigan 48105, USA

Accepted 3 April 2006

Diabetic distal symmetrical polyneuropathy (diabetic neuropathy, DPN) is a common late complication of chronic diabetes mellitus. Numerous clinical trials have sought to identify potentially effective treatment modalities for DPN. Endpoint measures for these studies have included symptom scores, neurologic examination scores, nerve conduction measures, quantitative sensory measures, and tests of autonomic function. Issues other than the pathophysiology of the disease include the appropriateness, clinical applicability, accuracy, and precision of the available measures of disease progression and response to therapy.

DPN is a slowly progressive disorder and thus small changes need to be detected over a prolonged study period. Reliable detection of progression is difficult, since changes in measurement of DPN may result from test-to-test variability rather than from change in disease or therapeutic intervention.

The San Antonio neuropathy consensus called for study designs requiring multiple, time-consuming, often expensive electrophysiologic, quantitative sensory, and clinical tools to document disease progression and response to therapy.[2,3] From the standpoint of clinical trial design, it is helpful to know the variability inherent in each measurement to determine the appropriate number of subjects necessary for the trial. In addition, the effect of using multiple centers is another important, yet unknown, component of measurement variability.

Zenarestat,[22] a highly potent aldose reductase inhibitor (ARI), was evaluated in a large, multicenter Phase 3 trial of mild DPN,[9] using guidelines provided by consensus panels.[2,3] A battery of nerve conduction studies (NCS) and quantitative sensory tests (QST) was performed in triplicate at each endpoint for each patient enrolled. This study was one of the largest long-term, placebo-controlled clinical trials investigating DPN. A significant increase in serum

Abbreviations: ARI, aldose reductase inhibitor; CRCC, Central Reading and Coordinating Center; CV, coefficient of variation; CDT, cool thermal detection threshold; DPN, diabetic peripheral neuropathy; EMG, electromyography; HbA_{1C}, glycosylated hemoglobin; ICC, intraclass correlation coefficients; IQR, interquartile range; JND, just noticeable difference; NCS, nerve conduction studies; PNSS, Penn Neuropathy Symptom Scale; QST, quantitative sensory testing; VDT, vibration detection threshold

Key words: aldose reductase inhibitors; diabetic neuropathy; nerve conduction studies; quantitative sensory testing; zenarestat

Correspondence to: S. Watling; e-mail: sharon.watling@pfizer.com

Published online 17 May 2006 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/mus.20577

creatinine observed in some zenarestat-treated patients resulted in an early termination of the pivotal study and the discontinuation of clinical development of zenarestat.

Despite early termination, sufficient NCS and QST data are available to report on the reproducibility, site-to-site variability, value of multiple assessments, learning potential, and impact of repetitive testing on sample size for NCS and QST at baseline and following 12 months of placebo or zenarestat therapy. The implications of these findings are discussed in the context of future DPN clinical trials.

SUBJECTS AND METHODS

Study Design. Detailed study design information is available in a previous publication.[9] In brief, this trial included over 1100 patients at 40 sites in the USA and Canada. Patients were screened by a trained nurse practitioner or physician using physical examination, medical history, and the Penn Neuropathy Symptom Scale (PNSS) to determine whether clinically mild neuropathy was present.[8] Patients screened were men or women aged 18–70 years with clinically stable type 1 or 2 diabetes mellitus for at least 6 months, glycosylated hemoglobin (HbA_{1C}) <12%, and stable/optimized antidiabetic therapy for at least 3 months. Patients with other neurologic disorders, other relevant diseases, significant laboratory abnormalities, and women who were pregnant, lactating, or of childbearing potential were excluded. The study was conducted according to the principles of the Declaration of Helsinki and approved by the ethics committee or institutional review board at all 40 study sites.

After the initial screening examination, the presence of mild distal symmetrical DPN was confirmed by a comprehensive neurologic examination administered by a board-certified neurologist. Baseline and subsequent NCS and QST data and waveforms were reviewed at the Central Reading and Coordinating Center (CRCC, Department of Neurology, University of Pennsylvania). Bilateral, recordable, and CRCC-confirmed sural sensory responses and a left median distal motor latency of <4.6 ms (to exclude moderate or severe carpal tunnel syndrome) were required for eligibility.

In addition to clinical evidence of DPN, at least one abnormal NCS or QST measurement was required for enrollment into the study.[9] Abnormal NCS was defined as 2.5 standard deviations below (velocity and amplitude) or above (latency) the mean for age (velocity, amplitude, and latency), height (velocity and latency), or body surface area (amplitude). Abnormal values were determined by taking into account the age, gender, height, and weight of the individual using the Neuropcentiles database software (WR Medical Electronics Co., Stillwater, Minnesota, based on results obtained elsewhere[15,25]).

Abnormal QST was defined as vibratory or cool thermal perception threshold 1.5 standard deviations above the mean for age. This level was selected because it was difficult to find patients with clinically established neuropathy as well as intact sural sensory responses who had QST values greater than 2.5 SD from the mean. Preliminary studies by us showed that the more strict QST criteria (>2.5 SD) led to such a high screening failure rate (due to the absence of sural sensory responses) that it would have precluded performance of the study.

Patients considered eligible were stratified by baseline HbA_{1C} ($\leq$ or >8%) and randomized to one of three zenarestat treatment groups (placebo, 600 mg/day, or 1200 mg/day). Adjustment of antidiabetic medications to achieve American Diabetes Association guidelines was allowed during the study.

Nerve Conduction and Quantitative Sensory Evaluations. The CRCC neurologists and technologists trained and certified all individuals who performed NCS and QST testing. This training included evaluations of the technical quality of normal tracings from each tester. Sites were required to pass a CRCC certification process prior to screening the patients. NCS tests were performed or supervised by a "certified" electromyographer at each site.

Certification was established by one of several requirements: board certification by the American Board of Electrodiagnostic Medicine, and added qualifications in clinical neurophysiology by the American Board of Psychiatry and Neurology, or the Canadian equivalent. In addition, each electromyographer was trained on the study equipment and the testing protocol at an investigators' meeting or at the CRCC and was required to perform CRCC-approved studies on normal individuals before patient studies were conducted. Those performing QST studies were trained on the equipment and study protocol by WR Medical (Stillwater, Minnesota). They were also required to perform protocol-specific QST studies on normal individuals and have them approved by the CRCC prior to conducting QST studies in study patients.

All NCS and QST data and waveforms were faxed to the CRCC. Each page was reviewed and approved by the CRCC prior to inclusion into the study database. Technically unsatisfactory studies were re-

peated. Each waveform and data sheet was individually reviewed by both a CRCC technologist and a neurologist. Of the 18,489 studies reviewed, 19% needed corrections and retests were required in 4%. These figures are comparable to the correction rate of 31% and rejection rate of 9% reported by Bril and colleagues.[6]

Nerve conduction studies were repeated in triplicate on separate days within a 4-week window. These were performed at baseline, month 12, and month 24, using a two-channel Viking Quest electromyography machine (Nicolet Biomedical, Inc., Madison, Wisconsin). Velocity (median forearm sensory, sural sensory, and peroneal motor), F-wave latency (median and peroneal), and amplitude (median and sural sensory nerve action potentials) were assessed on the left side of the subject unless that side could not be tested.

Standard NCS surface recording methods were used. Sural and median sensory responses were recorded antidromically. Median sensory forearm conduction velocity was calculated from the positive peak onset latencies after stimulation at the elbow and wrist to minimize the potential confounding effects of carpal tunnel syndrome. Peroneal motor conduction velocity was calculated using the onset latencies with stimulation at the ankle and fibular head. F-wave latencies were the minimum onset latency of at least 4 unequivocal F-wave responses obtained after a minimum of 16 stimulations. The techniques were standardized by the centralized training sessions and a study-specific methods manual was provided to each site. Near-nerve skin temperature was maintained at ≥32°C for the arm and ≥31°C for the leg.

The CRCC staff reviewed the NCS tracings with the temperatures printed on each page. They identified a small number (<0.5%) of studies that were below the minimum temperature and only a few individual recordings that were more than 2°C above the minimum temperature. In each case, the site was contacted and instructed to repeat the studies that fell outside of that range. As a result, the limbs were studied at 32°–34°C in the arm and 31°–33°C in the leg.

Quantitative sensory threshold testing was conducted in triplicate on 3 separate days within a 4-week window. These were performed at baseline, month 12, and month 24 using the CASE IV system (WR Medical Electronics Co.). The CASE IV system provides precise control over the stimulus intensity, mandatory equipment calibration verification routines, standardized patient instructions, and the use of computer-generated and validated algorithms.[13,16,23] Both vibration detection threshold (VDT, great toe) and cool thermal detection threshold (CDT, dorsal foot) on the left were assessed under controlled-temperature conditions. A 4-2-1 stepping algorithm was used for the VDT and CDT testing.[17] Thresholds were expressed as 1 of 25 just-noticeable-difference (JND) units that varied from 1 to 25.

Technicians and physicians were instructed during the central training sessions for NCS and QST not to refer to prior studies on patients when performing the subsequent studies. Additional measures of neurologic function and safety were performed, but are not the emphasis of this study.[9]

Statistical Methods. Due to early termination of the study only baseline and 12-month data are presented, as only insufficient 24-month data are available. All patients completing the 12-month assessment were included. Change from baseline in the treatment groups was analyzed using a paired *t*-test within the treatment group.

NCS or QST recordings that were considered technically nonevaluable by the CRCC staff were recorded as missing. Studies not performed were recorded as missing. Technically satisfactory tracings with undetectable responses were imputed as follows: nerve conduction velocity: the first percentile of the patients' data at baseline or 12 months for that nerve; sensory amplitude: 0 μV; F-wave latency: missing; and QST: 25 JND.

Imputed values are different for different variables because the physiologic attributes of each differ. Amplitude decreases through a linear scale to below 1 μV and then ultimately is absent. The reasonable imputed value is 0 μV. Conduction velocities differ; that is, they decrease to about 70% of the lower limit of normal velocity, below which the response is typically absent. The imputed value is then chosen to best estimate the velocity at the time that it was absent. QST testing with no sensation at all was imputed at the highest JND on that scale of 1–25. In any event, as noted in our previous work,[9] there were few imputed values. The sural response, generally the first to decline with axonal neuropathies, had the most imputed values. As the neuropathy worsened, some sural values were lost, but even this was observed in only 20 out of 355 patients.

Data presented in the previous study summarizing the trial results[9] and also in this study that provides sample size estimates for a range of replicates of the various parameters include the imputed values. These imputed values were included in these analyses since clinical studies typically have unmeasurable values. However, the remaining data pre-

sented herein do not include imputed data in order to evaluate the precision of the various electrophysiologic techniques.

Intersite variability in NCS and QST assessments at baseline was assessed using a random effects linear model that determined the proportion of variance due to site, patient, and random error. Total variance (the square of the standard deviation, SD) is presented to reflect that the statistical model partitions the variance, rather than the SD, into identified components (e.g., patient and site) and residual (random) error.

To assess the value of having replicate measures of NCS and QST, we defined three possible scenarios to compare treatment differences in the change from baseline to month 12 for the various NCS and QST measurements. Summary statistics were tabulated describing the change from baseline to month 12 for the first replicate, the average of the first two replicates, and the average of all three replicates for all parameters. Boxplots comparing the number of replicates, separated by treatment groups, were used to summarize graphically the effect of replicate measures.

Reproducibility of each parameter was assessed using intraclass correlation coefficients (ICCs)[10] among the three replicates at baseline. Coefficient of variation (CV) as a percentage for repeat testing was calculated in the standard manner (CV = 100 × standard deviation/mean) for all parameters using combined baseline data from all treatment groups.

To understand the trade-off between the number of replicates and sample size, we calculated sample sizes for a two-arm clinical trial with a continuous endpoint, such as change from baseline nerve conduction velocity, based on a standard *t*-test analysis. We calculated the sample size per treatment arm based on a two-sided 5% type I error rate and 90% power. For each parameter, the anticipated clinically meaningful drug effect was 1 unit (e.g., meters per second), with the exception of F-wave latency, which was 1.5 units (milliseconds) for the median motor nerve and 2.5 units (milliseconds) for the peroneal motor nerve. We chose these values as one estimate of the minimum change over 1 year that some would consider meaningful. Although there is no consensus about the smallest degree of electrophysiologic change that is clinically meaningful, this model, with the methods detailed in the Appendix, is designed to accommodate any values chosen as "clinically meaningful" to generate sample sizes.

We assumed that the variance, *V*, consists of two components: one due to differences between subjects, *S*, which is not reduced by taking replicate assessments, and one due to assessment variation, *A*, which is reduced by taking replicates. Estimates of variability for sample size calculation were based on the formula:

$$V = S + A/r$$

where r is the number of replicate assessments, and estimates of *S* and *A* were calculated from the zenarestat study. The model assumed equal variance at baseline and at the end of 12 months. Further details on this methodology are available in the Appendix and can be modified to accommodate specific definitions for clinically meaningful changes (e.g., 2 m/s rather than 1 m/s).

To assess whether fewer replications at month 12 were needed in comparison to the number of replicates at baseline, variance for each parameter was calculated at these two timepoints and compared using a variance components model.

RESULTS

The baseline demographics and electrophysiologic data from this zenarestat trial have been summarized in a previous publication.[9] Baseline data from this clinically defined mild neuropathic population show consistent characteristics across treatment groups. The overall population had a mean age of approximately 52 years, consisted primarily of patients with type 2 diabetes, had a mean baseline HbA_{1C} of 7.8%, and had a mean duration of diabetes of 10 years.[9]

The previous publication also depicts the change in electrophysiologic data over 12 months.[9] In the placebo group, nerve conduction velocities showed a discernible and, in some cases, statistically significant decline. Patients treated with zenarestat showed slowing of progression or improvement of neuropathy at 12 months as assessed by nerve conduction velocities. Quantitative sensory testing of cold thermal sensation in the placebo group showed worsening over 12 months. Vibratory testing was insensitive to changes over 12 months. In contrast to NCS, neither cool thermal nor vibratory testing was able to identify zenarestat treatment effects over a 12-month treatment period.

Variability due to testing at multiple study centers was small (Table 1). Patient variability accounted for the majority of the overall variance. This finding was consistent among all measured NCS and QST parameters.

For NCS, the standard deviation around the mean change from baseline decreased in a stepwise manner as the number of replicates increased from

Table 1. Proportion of measurement variance in baseline replicates.

Parameter	Number of replicates	Total variance*	Variability (% of total)		
			Site	Patient	Random error
Nerve conduction studies[†]					
Median forearm sensory	4276	23.72	1.47 (6%)	13.86 (58%)	8.33 (35%)
Peroneal motor	4275	23.61	1.27 (5%)	19.31 (82%)	3.23 (14%)
Sural sensory	4262	30.15	3.09 (10%)	20.14 (67%)	7.49 (25%)
Quantitative sensory testing					
Cool thermal	4136	20.88	1.15 (6%)	14.41 (69%)	5.58 (27%)
Vibration	4155	10.08	0.56 (6%)	7.79 (77%)	1.89 (19%)

*Variance = (standard deviation)^2.
[†]Nerve conduction studies listed are conduction velocities (m/s).

one to three. For example, with peroneal motor conduction velocity, the standard deviations in the placebo group for one, two, and three replicates were 3.1, 2.5, and 2.2, respectively. The standard deviations for one to three replicates for the sural sensory conduction velocity (4.5, 3.6, and 3.2), median forearm sensory conduction velocity (4.5, 3.6, and 3.2), median sensory amplitude (6.3, 5.3, and 4.5), median F-wave latency (1.3, 1.0, and 0.9), and peroneal F-wave latency (3.8, 3.4, and 3.1) showed similar improvement. The groups treated with zenarestat 600 mg and 1200 mg showed a similar pattern. The one exception to this pattern was sural sensory response amplitude in the placebo group (SD for one to three replicates: 3.1, 2.5, and 3.0, respectively).

Figures 1 and 2 display representative boxplots of QST parameters from all treatment groups. The standard deviations around the mean change from baseline decreased with increasing number of replicates in all QST cases. QST measures in the placebo group yielded the following pattern of standard de-

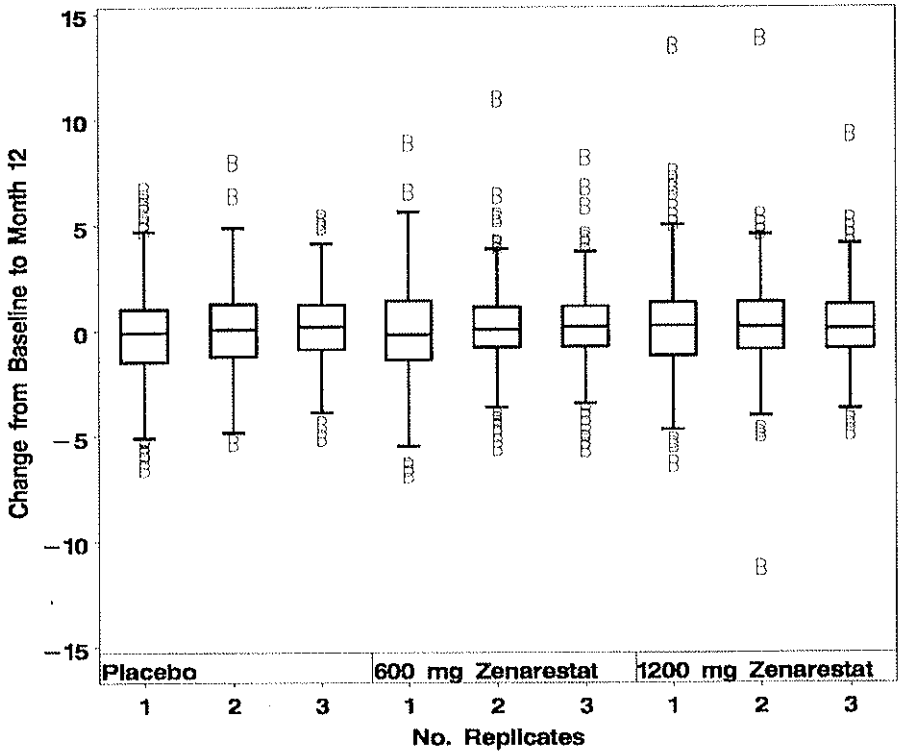

FIGURE 1. Analysis of QST variability: vibration. Box = 1st to 3rd quartiles (interquartile range), line in box = median, whiskers extend < 1.5*IQR.

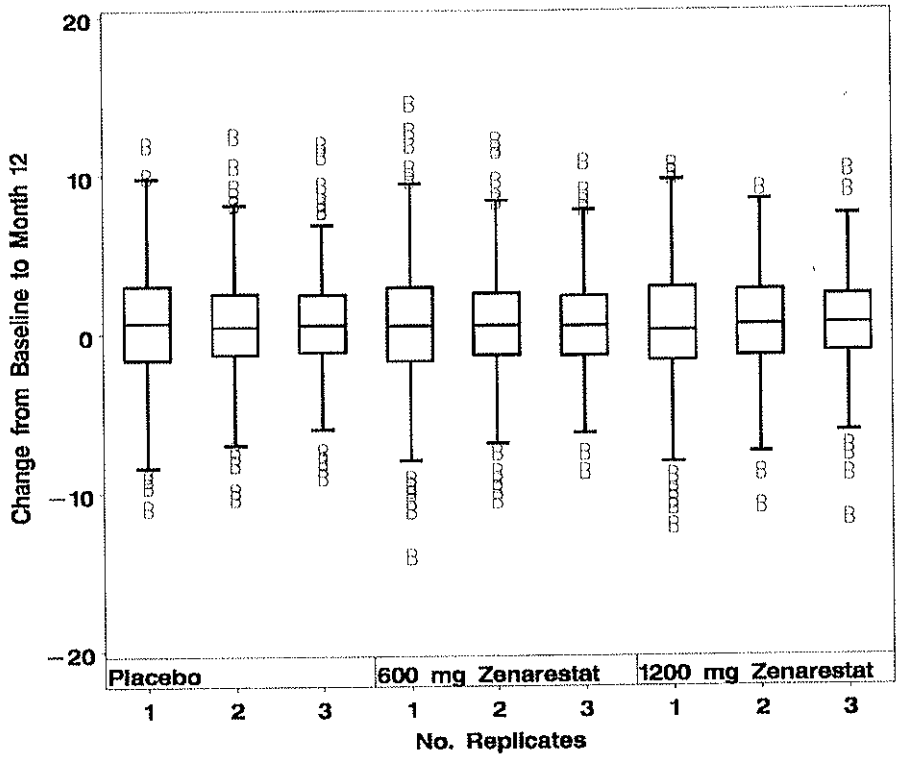

FIGURE 2. Analysis of QST variability: cool thermal. Box = 1st to 3rd quartiles (interquartile range), line in box = median, whiskers extend < 1.5*IQR.

viation values: vibration (2.3, 1.9, and 1.7) and cool thermal (3.9, 3.3, and 3.1) for one to three replicates, respectively.

The ICCs quantifying the degree of agreement among the replicates at baseline are shown in Table 2. In general, ICCs exceeded 0.6, indicating that nerve conduction studies were highly reproducible. Many ICCs exceeded 0.8, including peroneal motor NCV, median forearm sensory response amplitude, and both peroneal and median F-wave latencies. In the case of QST, the range of ICCs was also acceptable, exceeding 0.6 in all cases. Vibratory QST was most consistent, with ICCs ranging from 0.79 to 0.81 for the three treatment groups (Table 2).

Table 3 delineates the CVs for each parameter measured at baseline. By comparing the CVs for various parameters, the relative precision of a given

Table 2. Range of intraclass correlation coefficients for electrophysiologic and QST parameters at baseline for one to three replicates for all treatment groups.

Parameter	ICC range
Nerve conduction velocity	
Median forearm sensory	0.64–0.70
Peroneal motor	0.84–0.88
Sural sensory	0.73–0.77
Amplitude	
Median forearm sensory	0.91–0.92
Sural sensory	0.76–0.80
F-wave latency	
Median motor	0.87–0.90
Peroneal motor	0.83–0.89
Quantitative sensory testing	
Cool thermal	0.68–0.73
Vibration	0.79–0.81

Table 3. Coefficients of variation for three replicates at baseline for all treatment groups.

Parameter	Coefficient of variation (%)
Nerve conduction velocity	
Median forearm sensory	7.68
Peroneal motor	11.41
Sural sensory	11.89
Amplitude	
Median forearm sensory	52.32
Sural sensory	53.43
F-wave latency	
Median motor	8.38
Peroneal motor	10.77
Quantitative sensory testing	
Cool thermal	30.22
Vibration	14.87

Table 4. Effect of increasing the number of replicates on sample size per treatment arm, based on baseline variance.*

	Sample size*								
	Nerve conduction velocity			Amplitude		F-wave latency		Quantitative sensory testing	
Number of replicates	Peroneal motor	Median forearm sensory	Sural sensory	Median sensory	Sural sensory	Peroneal motor	Median motor	Cool thermal	Vibration
1	127	281	333	520	164	37	10	261	74
2	92	208	257	398	122	30	7	201	55
3	81	184	232	358	109	28	6	182	48
4	75	172	219	337	102	26	5	172	45
5	71	165	211	325	97	26	5	166	43

Based on combining all three treatment arms at baseline and assuming equal variance at baseline and 12 months.

parameter (its SD relative to its arithmetic mean) could be compared. For example, amplitude was much less precise than was NCV due to the wide variance between patients. Likewise, cool thermal QST was less precise than was vibratory QST.

Table 4 demonstrates the effect of repeated testing on sample size. The decreasing variability of each measure with increasing number of replicates lessened the sample size needed to show a "clinically meaningful" difference. Although additional replicates would continue to decrease the necessary sample size, this effect was a diminishing one and patients are unlikely to agree to undergo more testing sessions per endpoint. Therefore, data consistent with one to five replicates are shown.

In order to assess the possible existence of a learning curve, error (variance without the component of variability between subjects) for the various parameters was calculated at baseline and month 12 (Table 5). When comparing error at baseline in relation to error at 12 months, a small decrease did occur, except for sural sensory response amplitude. Several parameters (QST cool thermal, sural and median sensory response amplitude, and median F-wave latency) showed statistically significant differences between baseline and 12 months. This small decrease in error is unlikely to reduce the number of replicates needed at month 12 in comparison to the number required at baseline.

DISCUSSION

Nerve conduction studies are an important objective technique for quantifying peripheral neuropathy in diabetic neuropathy trials.[2,3] This study, as well as numerous others, demonstrates that, when performed with attention to detail, NCS are both reliable and reproducible. NCS also provide a more monotonic response than other quantitative mea-

Table 5. Variance of NCS and QST parameters at baseline and 12 months.

Parameter	Baseline error*	12-month error*	Subject
Nerve conduction velocity			
Peroneal motor	3.54 (24%)	3.09 (21%)	2.71 (55%)
Median forearm sensory	7.56 (23%)	6.20 (19%)	6.46 (58%)
Sural sensory	7.64 (19%)	6.86[†] (17%)	8.59 (64%)
Amplitude			
Median sensory	12.08 (19%)	11.04[†] (18%)	13.14 (63%)
Sural sensory	4.06 (21%)	3.88[†] (20%)	3.82 (59%)
F-wave latency			
Peroneal motor	5.21 (18%)	3.56 (12%)	6.58 (69%)
Median motor	0.83 (33%)	0.59[†] (24%)	0.36 (43%)
Quantitative sensory testing			
Cool thermal	6.59 (21%)	4.70[†] (15%)	6.75 (64%)
Vibration	1.92 (22%)	1.74 (20%)	1.67 (58%)

Where error is the component of variance due to factors other than variability between subjects.
[†]$P < 0.05$.

sures of diabetic neuropathy.[18] QST, the quantitative assessment of sensory thresholds, is not only a research tool, but also a measure of the actual clinically experienced sensory loss by the patient. QST measures will likely grow in importance in clinical trials of neuropathy because these deficits correlate well with foot ulcers and other deleterious effects of neuropathy.[1,5] Our data show that these studies can be performed with excellent reproducibility even when performed at many sites.

NCS and cool thermal QST were consistently able to detect deterioration of neuropathy in the placebo population over a 12-month period.[9] Vibratory QST was less able to show this deterioration, perhaps due to the relatively slower rate of decline in large-fiber activity reflected by vibratory testing[4] and the baseline variability among patients within the study population. The inability to detect deterioration is not due to inherent problems with test reproducibility.

Zenarestat treatment was associated with the slowing of disease progression and, in some cases, an improvement, as measured by NCS.[9] Treatment effects were not discernible with QST. The effects were more likely due to the relatively slow loss of sensory function with DPN rather than inherent test variability. Unfortunately, despite evidence of efficacy, adverse renal effects halted the development of this potent aldose reductase inhibitor. It is unknown whether QST assessments at 24 months would have yielded positive results as the number of patients assessed after 24 months of therapy was too small to provide convincing results.

The minimal contribution of intersite variability to overall variance (Table 1) has been shown.[21] This may reflect tight criteria and selection of patients, standardized methods of assessment, meticulous attention to training, certification, and reading of NCS and QST by a central reading center, and perhaps other reasons. The techniques of NCS training and data review by the central reading center used in this trial were very similar to those reported by Bril and colleagues.[6] This study extends that experience to QST as well. The good performance of QST may also lie, at least in part, with the CASE IV system, with its precise control over the stimulus intensity, mandatory equipment calibration verification routines, standardized patient instructions, and the use of computer-generated and validated algorithms.

NCS have been used in most recent clinical trials of diabetic neuropathy as they have been considered the least variable and most reliable, albeit surrogate, measure of nerve function. The reproducibility of the various NCS attributes in this study (Table 2) was better than most previously reported data from multicenter clinical trials (most done in duplicate at baseline),[5,26–28] but not quite as good as that reported in another study performed with triplicate measures.[6]

The use of QST to follow patients' sensation throughout the trial requires that the test be highly reproducible. QST studies with CASE IV have been highly reproducible when performed at a single site.[12,14] The lack of good data on the reproducibility of such measures with modern instruments at multiple sites has recently been reviewed.[12] To date, there have been limited data on the reproducibility of QST measures in multicenter clinical trials.[21,27,28] Our data show that measures of cool thermal and vibration thresholds with CASE IV instrumentation are highly reproducible (Table 2) and would suit the needs of such studies. QST variability in this multicenter study was better than that found in other multicenter trials.[7,14,21,24,27–29] A remarkable finding was that QST, as used in this study, had a reproducibility comparable to that of nerve conduction studies (Table 2).

Nerve conduction velocity and QST data at baseline showed significant intersubject variance. Intersubject variance was the most significant contributor at baseline and outweighed all other sources of variance. This wide range of electrophysiologic patient characteristics occurred despite using multiple clinical methods to define a homogeneous patient population with mild neuropathy. The present study incorporated screening tools, such as the Penn Neuropathy Score,[8] Michigan Neuropathy Scoring Instrument,[19] presence of bilateral sural sensory nerve action potentials, and clinical examination, to define a mild/moderate neuropathy population for entrance into the zenarestat trial. The baseline variance and significant difference among patients within the population not only increases the sample size needed to detect treatment effects but also raises concerns regarding the correlation between the clinical and the electrophysiologic definitions of mild to moderate neuropathy. This is a question of continued debate. The resolution of this issue is of key importance to future clinical trials and the development of therapies for neuropathy.

Triplicate measures of NCS and QST at each end-point were performed in the zenarestat trial despite the recognized time, effort, and cost incurred. Our data suggest that these additional replicates significantly decrease the number of patients required to show a statistically significant treatment effect. Although the values chosen as "clinically significant" are the subject of ongoing debate, the relationship still exists and details in the Appendix

show how to adjust the "clinically meaningful" criteria to calculate sample sizes for other definitions of clinical significance.

F-wave latencies are highly reproducible, a finding that translates to a small patient population needed to show a statistically significant change from baseline. Although statistically relevant, this finding may not be as clinically relevant, as more variable measures of distal nerve function more directly reflect the motor and sensory deficits of DPN.

The use of a composite rank score (using a combination of NCS and QST measures) may have led to a further reduction in measurement error and the allowance of a smaller number of patients to show a statistical difference. However, we chose to analyze the individual parameters that might be used to make up such a score. This may be particularly helpful in the case of the QST measures, because they reflect actual clinical deficits rather than composite scores that may represent less biologically meaningful data.

The design of future clinical trials will require weighing the cost of replicate measures against the logistics of recruiting and retaining patients. Patient recruitment in neuropathy trials is often difficult, so that repeating measures is often worth the additional measurement costs necessary to decrease sample size. Another consideration is patient acceptability. The continued repetition of tests beyond 3 or 4 replicates decreases sample size but would not likely be acceptable to patients or institutional review boards.

The basis of assessing the effect of replicates on the variability of measurements at baseline and 12 months was to determine whether a learning effect occurs at the site and patient level. One could envision decreased variability over time as a site gains more experience and patients become less anxious and more cooperative with these measures.

It is difficult to determine mathematically the amount of variance due to learning alone, even in the placebo population. Multiple factors and interventions occurred during the 12-month treatment period. The number of patients seen at a site and the timing of examinations in relation to the study period varied widely among sites due to recruitment rate. In many cases population sample size decreased over the study period due to multiple factors. Variance did decrease over time, but this decrease is difficult to attribute to a single intervention such as improved adherence to technique.

Data from this large, multicenter Phase 3 zenarestat trial have provided the opportunity to assess variability and reproducibility of multiple measures of nerve function. NCS provided reproducible data, which has been linked to clinical outcome measures.[11] In addition, this study showed that QST techniques are also highly reproducible, even when performed at multiple sites.

Repeating NCS and QST measures helps decrease the variability between subjects, leading to smaller sample sizes for clinical studies. This observation is of critical importance for the conducting and funding of neuropathy research. Consideration of these data, particularly with regard to QST, should aid in the design of future clinical trials investigating DPN.

THE ZENARESTAT STUDY GROUP

The following persons participated in the 24-month, double-blind, randomized, placebo-controlled, fixed-dose, parallel-group Multicenter Study of Zenarestat in the Treatment of Diabetic Neuropathy Trial: Central Reading and Coordinating Center: S. Bird, M. Brown (Philadelphia, Pennsylvania). Principal Investigators: D. Zochodne (Calgary, AB, Canada); D. Studney, C. Kreiger (Vancouver, BC, Canada); R. A. Kaplan, R. Stevens (Concord, California); W. Feng, N. Slatkin, M. B. Davidson, J. Nadler (Duarte, California); S. Edelman, G. Shehan (San Diego, California); R. Olney, A. Poncelet (San Francisco, California); R. J. McCarthy (San Rafael, California); R. Winer, A. Starr, A. Charles, J. See (Tustin, California); R. L. Weinstein, R. Stevens (Walnut Creek, California); J. Goldstein, S. Novella (New Haven, Connecticut); A. Berger (Jacksonville, Florida); P. N. Weissman, B. Aiken, E. Carrazana, V. Faradji (Miami, Florida); J. Glass (Atlanta, Georgia); R. F. Arakaki, M. Yee, D. Kaku (Ewa Beach, Hawaii); M. S. Kirkman, J. Kincaid, B. Gumbiner (Indianapolis, Indiana); V. Fonseca, M. Shamsnia (New Orleans, Louisiana); E. Feldman, D. A. Greene, J. Russell (Ann Arbor, Michigan); G. Grunberger, R. Lewis, J. Selwa (Detroit, Michigan); P. Kelkar, G. Parry (Minneapolis, Minnesota); S. H. Horowitz (Columbia, Missouri); C. Walden (Richmond Heights, Missouri); J. R. Storey (Albany, New York); K. Hershon, E. Condon, M. Vishnubhakat (New Hyde Park, New York); H. Lesser (Rochester, New York); J. M. Shefner, C. S. Calder (Syracuse, New York); J. Buse, J. F. Howard (Durham North Carolina); V. Bril (Toronto, ON, Canada); L. Olansky, M. Trebbey (Oklahoma City, Oklahoma); A. McCall, Y. So, W. Johnston (Portland, Oregon); M. J. Guiliani, D. A. Kelley (Pittsburgh, Pennsylvania); A. Belanger, M. J. Monette, E. LaLumiere (Laval, QC, Canada); T. Lin, D. Redmond, T. Hwang (Columbia, South Carolina); S. Aronoff, M. Vengrow (Dallas, Texas); P. Raskin, H. Unwin (Dallas, Texas); A. J. Garber, J. M. Killian (Houston, Texas); S. L. Schwartz, M. Merren (San Antonio, Texas); M. Bromberg (Salt Lake City, Utah); E. C. Yuen (Seattle, Washington).

This study was supported by Parke-Davis Pharmaceutical Research, now Pfizer, Inc., New York, New York.

APPENDIX: ASSESSING THE BALANCE BETWEEN THE NUMBER OF SUBJECTS AND THE NUMBER OF REPLICATES

In general, increasing the number of replicates per subject decreases variability and thus the sample size in a clinical trial, but there is a point of diminishing returns. To better understand the relationship between the number of replicates and number of sub-

jects, we used a standard (two-sample *t*-test) sample size formula[20] for the simplest study design, a two-armed, placebo-controlled study with a continuous outcome as the primary endpoint (e.g., change from baseline in NCV or QST):

$$N = 2V\,(z_{\alpha/2}+z_{\beta})^2/\Delta^2$$

where N is the number of subjects per treatment arm; V is the variance of the endpoint, here estimated from the zenarestat study; α is the proscribed false-positive rate, here set at the two-sided 5% level; β is the proscribed maximum acceptable false-negative rate, here set to 10% (alternatively, characterized as 90% power); Δ is the hypothesized drug effect, for which we used a value of 1 unit for all outcomes, with the exception of 1.5 units for median motor F-wave latency and 2.5 units for peroneal motor F-wave; and z_p is the $100 \times (1 - p)$th percentile of the standard normal distribution (where $p = \alpha/2$ or $p = \beta$ in the equation above).

To estimate the variability in each endpoint from our study, we assumed that the "error" variance, V, consisted of two components: one component (S) that was due to differences between the subjects (and that would therefore not be reduced by taking replicate assessments on a single subject) and another component (A) that was due to assessment variation (and this component would be reduced by replicated assessments). Thus, the model for variability is:

$$V = S + A/r$$

where r is the number of replicate measurements taken at baseline and at endpoint.

The estimates of variability that were used to produce Table 5 in the text are available from the corresponding author.

Using the information just given, one can calculate the required sample size for new studies using different assumptions based on the magnitude of a clinically significant change. For example, if one wished to detect a 2-m/s treatment difference with 80% power for a study investigating cool thermal QST as its primary endpoint, the required sample size per treatment arm would be 44, 34, 31, 29, and 28, respectively, for one, two, three, four, and five replicates (assuming the other design characteristics were similar to those above).

REFERENCES

1. Ali Z, Carroll M, Robertson KP, Fowler CJ. The extent of small fibre sensory neuropathy in diabetics with plantar foot ulceration. J Neurol Neurosurg Psychiatry 1989;52:94–98.
2. American Diabetes Association, American Academy of Neurology. Proceedings of a consensus development conference on standardized measures in diabetic neuropathy. Neurology 1991;42:1823–1839.
3. American Diabetes Association, American Academy of Neurology. Report and recommendations of the San Antonio conference on diabetic neuropathy. Diabetes 1988;37:1000–1004.
4. Bird SJ, Brown MJ. Diabetic neuropathies. In: Katirji B, Kaminski H, Preston D, Ruff R, Shapiro B, editors. Neuromuscular disorders in clinical practice. Boston: Butterworth-Heinemann; 2002. p 598–621.
5. Boulton AJM, Hardisty CA, Betts RP, Franks CI, Worth RC, Ward JD, et al. Dynamic foot pressure and other studies as diagnostic and management aids in diabetic neuropathy. Diabetes Care 1983;6:26–33.
6. Bril V, Ellison R, Ngo M, Bergstrom B, Raynard D, Gin H, Roche Neuropathy Study Group. Electrophysiological monitoring in clinical trials. Muscle Nerve 1998;21:1368–1373.
7. Bril V, Kojic J, Ngo M, Clark K. Comparison of a neurothesiometer and vibratron in measuring vibration perception thresholds and relationship to nerve conduction studies. Diabetes Care 1997;20:1360–1363.
8. Brown MJ, Bird SJ. A simple diabetes neuropathy screening scale can predict measurable sural sensory responses. Muscle Nerve 1998;21:1576s.
9. Brown MJ, Bird SJ, Watling S, Kaleta H, Hayes L, Eckert S, et al. Natural progression of diabetic peripheral neuropathy (DPN) in the zenarestat study population. Diabetes Care 2004;27:1153–1159.
10. Cappelleri JC, Ting, N. A modified large-sample approach to approximate interval estimation for a particular intraclass correlation coefficient. Stat Med 2003;22:1861–1877.
11. Carrington AL, Shaw JE, Van Schie CH, Abbott CA, Vileikyte L, Boulton AJ. Can motor nerve conduction velocity predict foot problems in diabetic subjects over a 6-year outcome period? Diabetes Care 2002;25:2010–2015.
12. Chong PST, Cros DP. Technology literature review: quantitative sensory testing. Muscle Nerve 2004;29:734–737.
13. Dyck PJ, Bushek W, Spring EM, Karnes JL, Litchy WJ, O'Brien PC, et al. Vibratory and cooling detection thresholds compared with other tests in diagnosing and staging diabetic neuropathy. Diabetes Care 1987;10:432–440.
14. Dyck PJ, Kratz KM, Lehman JL, Karnes JL, Melton LJ, O'Brien PC, et al. The Rochester diabetic neuropathy study: design, criteria for types of neuropathy, selection bias, and reproducibility of neuropathic tests. Neurology 1991;41:799–807.
15. Dyck PJ, Litchy WJ, Lehman KA, Hokanson BA, Low PA, O'Brien PC. Variables influencing neuropathic endpoints: the Rochester diabetic neuropathy study of healthy subjects. Neurology 1995;45:1115–1121.
16. Dyck PJ, O'Brien PC, Johnson DM, Klein CJ, Dyck PJB. Quantitative sensation testing. In: Dyck PJ, Thomas PK, editors. Peripheral neuropathy, 4th ed. Philadelphia: Elsevier; 2005. p 1063–1093.
17. Dyck PJ, O'Brien PC, Kosanke JL, Gillen DA, Karnes JL. A 4, 2 and 1 stepping algorithm for quick and accurate estimation of cutaneous sensation. Neurology 1993;43:1508–1512.
18. Dyck PJ, O'Brien PC, Litchy WJ, Harper CM, Klein CJ, Dyck PJB. Monotonicity of nerve tests in diabetes. Diabetes Care 2005;28:2192–2200.
19. Feldman EL, Stevens MJ, Thomas PK, Brown MB, Canal N, Greene DA. A practical two-step quantitative clinical and electrophysiological assessment for the diagnosis and staging of diabetic neuropathy. Diabetes Care 1994;17:1281–1289.
20. Friedman LM, Furberg CD, DeMets D. Fundamentals of clinical trials, 2nd ed. Littleton, MA: PSG; 1985. p 96.
21. Gelber DA, Pfeifer MA, Broadstone VL, Munster EW, Peterson M, Arezzo JC, et al. Components of variance for vibratory and thermal threshold testing in normal and diabetic subjects. J Diabetes Complic 1995;9:170–176.

22. Greene DA, Arezzo JC, Brown MB, Zenarestat Study Group. Effect of aldose reductase inhibition on nerve conduction and morphometry in diabetic neuropathy. Neurology 1999; 53:580–591.
23. Gruener G, Dyck PJ. Quantitative sensory testing: methodology, applications, and future directions. J Clin Neurophysiol 1994;11:568–583.
24. Luft D, Ziegler D. Evaluation of drug effects. In: Gries FA, Low PA, Cameron NE, Ziegler D, editors. Textbook of diabetic neuropathy. New York: Thieme; 2003. p 313–360.
25. O'Brien PC, Dyck PJ. Procedures for setting normal values. Neurology 1995;45:17–23.
26. Santiago JV, Sonksen PH, Boulton AJM, Macleod A, Beg M, Bochenek W, et al. Withdrawal of the aldose reductase inhibitor tolrestat in patients with diabetic neuropathy: effect on nerve function. J Diabetes Complic 1993;7:170–178.
27. Sundkvist G, Armstrong FM, Bradbury JE, Chaplin C, Ellis SH, Owens DR, et al. Peripheral and autonomic nerve function in 259 diabetic patients with peripheral neuropathy treated with ponalrestat (an aldose reductase inhibitor) or placebo for 18 months. J Diabetes Complic 1992;6:123–130.
28. Valensi P, Attali JR, Gagant S, and the French Group for Research and Study of Diabetic Neuropathy. Reproducibility of parameters for assessment of diabetic neuropathy. Diabetic Med 1993;10:933–939.
29. Vinik AI, Suwanwalaikorn S, Stansberry KB, Holland MT, McNitt PM, Colen LE. Quantitative measurement of cutaneous perception in diabetic neuropathy. Muscle Nerve 1995; 18:574–584.