Psychopharmacology (1997) 132 : 61–66

ORIGINAL INVESTIGATION

Michael P. Caligiuri · James B. Lohr
John Rotrosen · Lenard Adler · Philip Lavori
Robert Edson · Kathlene Tracy
Veterans Affairs Cooperative Study #394 Study Group

Reliability of an instrumental assessment of tardive dyskinesia: results from VA Cooperative Study #394

Received: 19 September 1996/Final version: 7 February 1997

Abstract Nine VA Medical Centers are participating in a 2-year double-blind placebo controlled study of antioxidant treatment for tardive dyskinesia (TD) conducted by the Department of Veteran Affairs Cooperative Studies Program. One of the principal outcome measures of this study is the score derived from the instrumental assessment of upper extremity dyskinesia. Dyskinetic hand movements are quantified by assessing the variability associated with steady-state isometric force generated by the patient. In the present report, we describe the training procedures and results of a multi-center reliability assessment of this procedure. Data from nine study centers comprising 45 individual patients with six trials each (three from left hand and three from right hand) were reanalyzed by an independent investigator and the results were subjected to reliability assessment. For the statistic of interest (average coefficient of variation over trials 2 and 3 for each hand, then take the larger of these two values), we found very high intraclass correlation coefficients for reliability over all patients across sites (ICC = 0.995). We also calculated the reliability of the measures across trials within patient for each combination of hand (right, left, dominant), rater group (site, control), and trials set (all three, trials 2 and 3). For a given hand and trial set, the reliability of the site raters was similar to that of the control. This study demonstrates that instrumental measures for the assessment of dyskinesia are reliable and can be implemented in multi-center studies with minimal training.

Key words Tardive dyskinesia · Instrumentation · Multi-center study · Antioxidant · Interrater reliability

M.P. Caligiuri (✉)
Motor Function Laboratory, Geriatric Psychiatry Clinical Research Center, V-116A1, San Diego VA Medical Center, San Diego, CA 92161, USA

J.B. Lohr
Psychiatry Service (116A), San Diego VA Medical Center, San Diego, CA 92161, USA

J.Rotrosen · L. Adler · K. Tracy
Psychiatry Service (116A), VA Medical Center, 423 East 23rd Street, New York, NY 10010, USA

P. Lavori · R. Edson
VA Palo Alto Health Care System, CSPCC (151K), 3801 Miranda Avenue, Palo Alto, CA 94304, USA

Introduction

The reliability, or consistency between measurements made under uniform conditions, is a necessary attribute for any assessment tool. Satisfactory levels of test-retest consistency have been achieved for several TD rating scales including the Abnormal Involuntary Movement Scale (Lane et al. 1985; Sweet et al. 1993) and the Simpson Tardive Dyskinesia Rating Scale (Simpson et al. 1979); however, the interrater variability often exceeds intra-rater variability for these scales (Bergen et al. 1984, 1988). Variation in the assessment of TD is an important clinical and research problem and has been linked to the inconsistencies in therapeutic responses to different agents (Richardson et al. 1982). Sources of variation may range from examiner bias to active suppression by the patient, to fluctuating biochemistry (Caligiuri et al. 1995).

Instrumental assessments have emerged to enhance our understanding of TD and facilitate our ability to prevent and treat it (Gardos et al. 1977). The Task Force Report on Tardive Dyskinesia (Kane et al. 1992) suggested that instrumentation may overcome some of the problems associated with observer ratings such as variable reliability, nonlinearity, and poor sensitivity. While laboratory instrumental assessments of TD have been available for over 25 years (see Lohr and Caligiuri 1992 for review), applications for the study of treatment outcome of TD have not achieved similar levels of acceptance in clinical practice. Lack of sufficient data

demonstrating the reliability of a particular instrumental method may be one important reason for this lapse. Previous laboratory studies have demonstrated relatively high test-retest reliability using instrumental techniques. Isometric force procedures have yielded intra-patient reliability coefficients of 0.85 (Caligiuri and Lohr 1990). Procedures which count movements such as position sensors (Trzepac and Webb 1987) or ultrasonic techniques (Resek et al. 1981; Bartzokis et al. 1989) are highly reliable with test-retest correlation coefficients greater than 0.90. Gattaz and Buchel (1993) reported an intra-subject reliability coefficient of 0.84, using an automated digital video image processing system which counted oral movements.

While instrumental procedures demonstrate high intra-subject reliability, little is known about their reliability assessed between multiple testers or across multiple centers or across multiple trials. One of the advantages of a multi-center study is that each participating center represents an independent replication of the study design, thus facilitating an examination of the consistency of the measurements (Meinert 1986). The Department of Veterans Affairs Cooperative Studies Program is conducting a multi-center study of antioxidant treatment for TD (CSP #394: "Vitamin E Treatment of Tardive Dyskinesia"). Instrumental procedures are included in the armamentarium for assessing treatment outcome, providing the first opportunity to assess multi-center reliability. The general aim of this paper is to report the reliability of an instrumental procedure across the nine study sites and to identify potential sources of between-examiner variability.

Materials and methods

Participating study sites

The following nine VA Medical Centers participated in this study: New York, N.Y.; Northport, N.Y.; West Haven, Conn.; Philadelphia, Pa.; Augusta, Ga.; Tucson, Ariz.; Portland, Ore.; West Los Angeles (Brentwood), Calif.; and San Diego, Calif. The first five patients from each site with baseline assessments were considered for the present study. This sample consisted of patients with a wide range of TD severity. The group baseline mean AIMS total score (items 1–7) was 10.8 (with a standard deviation of 4.2 and a range of 3–24). The mean AIMS score for the upper extremity item was 2.0 (with a standard deviation of 1.0 and a range of 0–4).

Instrumental procedures

The instrumental procedure quantified the error associated with the steady-state maintenance of isometric force by the hand. All assessments were performed blind to the treatment status of the patient. The force transducer consisted of a stainless steel rigid beam (10.5 cm × 1.8 cm × 0.4 cm) mounted onto a wooden platform (15 cm × 8.8 cm × 1.8 cm). The beam was instrumented with a pair of thin film resistive strain gauges (MicroMeasurements Group; Raleigh, N.C., USA) wired to form two active arms of a balanced four-arm Wheatstone bridge circuit. Pressure applied to the long axis of the beam deforms the strain gauge, thus changing the resistance of the current flowing through the bridge circuit. Applied pressure imparts a linear and proportional change in the transducer's output voltage which is sampled by the computer. Static and dynamic calibration of the transducer yielded a flat frequency response out to 40 Hz and a linear transfer function across a wide range of applied weights. The continuous voltage is calibrated in units of force (centiNewtons). The force transducer is capable of detecting changes in applied pressure of approximately ±2 cN. It is not likely that finger displacements caused by muscle contractions of this magnitude are detectable visually by the examiner.

Patients were instructed to exert constant low-level (300 cN of force) isometric contraction by keeping the hand in direct contact with the instrumented beam and to maintain that level by visually monitoring performance on the computer screen for 30 s. The finger flexor muscles (flexor digitorum profundus and the flexor digitorum superficialis) continuously contract to produce a near-constant level of force against the rigid beam. A dyskinetic patient will produce an irregular pattern of muscle contraction which is transduced as variable levels of force over time. Because the patient maintains a non-zero level of force around which the contractile force of the hand muscles are detected, contractile variations in both extensor and flexor muscles are detected. The procedure has been validated against standard clinical ratings (Caligiuri and Lohr 1990; Caligiuri et al. 1995).

Three 30-s trials per hand were conducted with the first trial serving as practice and to familiarize the patient with the instrumentation. For a given trial, a 5-s segment containing the largest variation in force was selected. The segment was low pass filtered using a bidirectional eight-pole digital Butterworth filter with a frequency cut-off of 3 Hz to remove instability due to parkinsonian tremor. For each trial, the mean acquired force and standard deviation of the 5-s filtered segment were obtained and used to calculate the coefficient of variation (CV), which is the standard deviation divided by the mean. The principal outcome for measurements performed at the sites was then obtained by calculating the average CV over trials 2 and 3 for each hand, and taking the larger of these two values.

The original data files were available for reliability assessment by the first author to provide control measures of force instability. Reanalysis was conducted independently from and used the same procedures followed by the sites. The principal outcome measure from the control rater was the average CV over trials 2 and 3 of his ratings for the hand with the larger average CV from the site ratings. If the average CV's from the site ratings were the same for both hands, we arbitrarily used the dominant hand to determine the control rater's outcome.

Study site training

The first author travelled to each study site over a period of 1 month to install the instrumentation and conduct training. Training involved approximately 2 h of demonstration with participation by the technician. Several data sets were acquired and analyzed over this time period. Within the first month of the study period, research assistants from each study site tested three patients (recruited for training purposes) and forwarded the data to the first author for inspection of the force waveforms. The purpose of this activity was to ensure that the data were being collected correctly with respect to signal offsets, trial sequence, sample rates, window duration and channel assignment. These data were reviewed and problems or inconsistencies were discussed at the kick-off meeting held in San Diego prior to the start of the study.

Guidelines were made available to each site to reduce inconsistency associated with problematic trials. These guidelines, described in Table 1, pertain to identifying the starting and ending points of each waveform and to removing spurious peaks.

Table 1 Guidelines used to reduce ambiguity over the location of starting and ending indices of the force waveform segments selected for analysis

Condition	Guideline
Starting index	Begin segment once waveform amplitude reaches target and remains on or near target for 2–3 s. Exclude initial segments suggestive of attempts to locate target
Ending index	End segment prior to drop in force which continues to baseline rest level. Ignore increases in force that occur just prior to end of trial
Spurious peaks or valleys that may be either movement related or caused by electrical artifact or noise	If a force peak (or valley) occurs in the absence of peaks (or valleys) having similar amplitudes on any other trial and if there is no documentation explaining the presence of unusual force peaks (or valleys) as due to movements unrelated to the task, place initial index to the right of the peak (or valley) if it occurs early in the trial or place the ending index to the left of the peak (or valley) if it occurs late in the trial

Statistical analyses of reliability

Data from five patients from each center (total of 45 patients) were available for reliability assessment. For 44 patients, there were 12 scores provided (six hand trial scores each from the site and control raters). One patient did not complete the procedure for the left hand, so the total number of data points was 534 ($44 \times 12 + 6$). These data were then sent to the study's statistician (R.E.) in Palo Alto for assessment of reliability using S-Plus software (S-Plus, 1993).

Of the many forms of reliability that have been described in the literature (Meinert 1986; Kramer 1988; Isaac and Michael 1990), we were concerned primarily with consistency, which refers to the reliability associated with the assessment procedure (Isaac and Michael 1990). The methodology described by Fleiss (1986), for when the same raters are involved in both the preliminary and main study, was used to determine the reliability among the data collected by the two sets of raters (site raters, and the control rater). Edson et al. (1997) summarize the assumptions associated with this methodology.

Fleiss (1986) discusses two options for estimating the interrater reliability (IRR) using the intraclass correlation coefficient (ICC). The conservative estimator involves a measure of variation among the rater sets; the other estimator does not, since it is assumed that the effect for each rater is set to zero or differences in the rater set effects can be adjusted for via stratification. The choice of ICC depends on the design of the study for which IRR is being estimated. A test of the hypotheses (by site and over all sites) that the effects for the two rater sets were both zero applied to the $45 \times 3 \times 2$ table indicated that the rater effects did not differ significantly, and were actually quite close to zero. In addition, the control rater's assessments are considered to be the "gold standard" for each patient. Thus, we decided to use the ICC that does not include a term for the rater variation.

If PMS and EMS are, respectively, the mean sum of squares for the subject and error terms resulting from application of standard analysis of variance (ANOVA) procedures to the data, the sample estimator of the ICC is:

$$\hat{R} = \frac{PMS - EMS}{PMS + (k-1)\,EMS}.$$

To construct a two-sided 95% confidence interval for $\hat{R}$, we used the bias corrected, accelerated non-parametric methodology (Efron 1987). The S-Plus code for this method (denoted bcanon) is contained in Efron and Tibshirani (1993). The bootstrap methodology involves the selection of B (in this case 2000) independent, with replacement random samples of size n from the n subjects. Since a small n will produce a confidence interval which is too wide for any practical purposes, we limited application of the bootstrap to the comparison of the two rater groups (site raters versus the control rater) over all sites.

The procedure used to collect the force instability data is almost totally automated; the only step not completed by the equipment is the selection of the 5-s interval with the largest variation in force. Thus, in essence the resulting estimated ICC indicates how similarly the site raters selected this interval compared to the control rater. Perhaps a more interesting result is to determine reliability of the force instability data across the three trials. To see if inclusion of the data from trial 1 affected reliability, separately for the six combinations of rater set (site, control) and hand (left, right, and dominant), we calculated estimated ICCs for two sets of trials: all three trials, and trials 2–3. Since our hypothesis was that inclusion of the trial 1 data would lower reliability, i.e., that there would be differences between the data across trials, we used the conservative version of the ICC. Using the notation above the sample estimate of the "conservative" ICC is:

$$\hat{R} = \frac{n(PMS - EMS)}{n\,PMS + (k-1)TMS + (n-1)(k-1)EMS}$$

where TMS is the mean sum of squares for the trial set term in the ANOVA. Confidence intervals were calculated using the bootstrap method described above.

Results

Two sets of results are presented. In the first set, we report the overall ICCs across sites for the principal outcome of the study which is the higher CV between the left and right hands with trials 2 and 3 averaged to form one CV. In the second set, we report the ICCs across trials to examine consistency between the first trial and the latter two trials. In both analyses, data from each of the nine study sites as well as the control site will be presented.

Table 2 summarizes the results of our analysis of the principal study outcome. We identify study sites by letter to avoid disclosure of low IRR. For the overall sample and for most of the sites, the mean and standard deviation for the assessments made by the site raters are almost equal to those of the control rater.

This similarity is reflected in the extremely high estimated ICCs for seven sites and for the overall sample, and the moderately high ICC estimate for sites C and E. The bootstrap 95% confidence interval on the ICC for the overall sample (0.988–0.998) is very tight around the point estimate of 0.995, and provides more evidence of the high reliability in the force instability measures. Close inspection of the data revealed that the sites with the lowest ICCs had the smallest range of scores with SDs of less than 0.50. Edson et al. (1997) discuss how a wider range in the scores produces a higher estimate of the ICC.

Table 3 summarizes the findings of the reliability estimates across the three trials for the left, right, and dominant hands and the 95% CIs on the ICCs by rater, hand, trial set. Point estimates of reliability were higher for the right hand than for the left, and for trials 2–3 than for trials 1–3. While not statistically different, the ICCs for trials 1–3 were much lower than the ICC for trials 2–3 for each hand as well as the dominant hand. These findings reflect less consistency between the first and subsequent trials and suggests a potential effect of learning which has not been previously reported.

Discussion

The present study represents the first attempt to examine the multi-center reliability of a relatively recent procedure. There have been only a few reports on the reliability of established assessment procedures in a multi-center study design. Since 1990, there have been reports on instrument reliability for multi-center studies of outcomes in cardiac surgery (Henderson et al. 1995), addiction research (Del Boca et al. 1994), dystonia (Defazio et al. 1994), outcomes of treatment for acute ischemic stroke (Albanese et al. 1994), and quantitative neuropathology assessment (Mirra et al. 1994) and cognitive changes (Kim et al. 1994) in Alzheimer's disease. In general, the findings from these studies suggested that the various aspects of reliability, including test-retest and interrater reliability, were favorable. The present finding that an electromechanical assessment procedure for tardive dyskinesia demonstrated excellent interrater reliability across nine study sites strengthens the clinical value of this procedure. The findings suggest that inexperienced raters can produce data that are about as reliable as those obtained from an experienced rater.

We found the instrumental procedure to be highly reliable across multiple study sites. The overall ICC from 45 patients was 0.995. This ICC from instrumental assessment of upper extremity TD was somewhat higher than the ICCs reported from observer ratings of TD. Smith et al. (1979) obtained a correlation (Pearson's *r*) of 0.76 for the upper extremity item of the AIMS and 0.87 for the total AIMS score in a study of approximately 300 patients. Lane et al. (1985) obtained an ICC of 0.65 from the same AIMS item and 0.79 for the total AIMS score in a study of 33 patients. Sweet et al. (1993) obtained and ICC of 0.44 for the upper extremity and 0.91 for total AIMS with ten elderly patients. We recently reported an ICC of 0.60 for upper extremity TD and 0.75 for the total AIMS, as part of the ongoing cooperative study involving ten patients and 18 raters (Edson, et al. 1997).

The discrepancy in ICCs between instrumental procedures and observer ratings of TD severity may be attributed largely to examiner-related factors; however, instrument-related factors may also play an important role. Vulnerability of multi-item rating scales such as the AIMS to the experience of the examiner have been discussed previously by Gardos et al. (1977), who stressed the importance of extensive training. Instrumental procedures, on the other hand, require minimal training as previously discussed in the

Table 2 Means and standard deviations (*SD*) for the five patients from each of nine study sites (labeled *A–I*) for the site rater and independent rater (control) and the intraclass correlation coefficient (*ICC*) for each site. Means and SDS are percentages derived by dividing the waveform SD by the waveform mean

Site	Site mean	SD	Control mean	SD	ICC
A	3.51	2.16	3.42	2.05	0.997
B	5.95	4.47	5.93	4.39	0.998
C	1.98	0.30	2.19	0.51	0.822
D	2.64	0.95	2.56	1.10	0.939
E	2.57	0.45	2.45	0.33	0.752
F	2.97	1.49	2.98	1.50	1.000
G	2.30	0.62	2.41	0.64	0.973
H	4.00	2.98	4.15	3.25	0.995
I	4.55	4.36	4.50	4.14	0.998
Overall	3.39	2.57	3.40	2.54	0.995

Table 3 Reliability estimates across the three trials for the left, right, and dominant hands (and the 95% confidence intervals on the ICCs) by rater, hand, and trial group

	Trials	Left hand	Right hand	Dom hand
Across sites	1–3	0.582 (0.326, 0.717)	0.821 (0.610, 0.929)	0.721 (0.375, 0.843)
	2–3	0.675 (0.561, 0.804)	0.924 (0.761, 0.975)	0.824 (0.585, 0.922)
Within control	1–3	0.652 (0.320, 0.887)	0.839 (0.567, 0.943)	0.711 (0.351, 0.881)
	2–3	0.767 (0.590, 0.913)	0.898 (0.673, 0.966)	0.797 (0.510, 0.928)

Methods section of this paper. Moreover, reliable observer ratings require a standard administrative procedure and clear instructions to raters regarding the procedure. Often these instructions are subject to individual interpretation which may compromise reliability. A minimal set of instructions are desirable as noted in Table 1 for the instrumental procedure. One of the features of an observer rating scale which may have an important influence on reliability is the dependence on anchor or threshold scores representing the extremes of the movement disorder. With observer ratings, a designation of "severe" is essentially arbitrary and based on the experience of the examiner, whereas with instrumentation, the criteria for severity is made explicit and is based on a continuous score.

The findings seem to support our decision to use trial 1 as practice. The results showing higher reliability for dominant compared with the nondominant hand bring to question the role of hand dominance in facilitating motor learning or in reducing short-term fluctuation. Further studies are needed to examine the influence of handedness on motor learning in patients with TD.

The generalizability and clinical usefulness of findings stemming from multi-center treatment studies such as CSP#394 depend largely upon the reliability of the outcome measures. The present study demonstrated that instrumental procedures for assessing the severity of upper extremity dyskinesia are reliable and with minimal training can be applied to multi-center clinical trials.

The advantages of the present device for assessing TD over other instrumental procedures have been discussed in a previously published review of instrumental assessments (Caligiuri 1997). In general, the choice of system depends primarily upon the objective of the study and secondarily upon the nature of the patient sample. For example, if an individual is interested in studying uncooperative or difficult-to-test patients with observable TD, accelerometry may be chosen because activation of this device does not require the cooperation of the patient. If, however, the purpose is to detect subtle change in the dyskinetic condition, or to minimize voluntary suppression of the dyskinesia, force steadiness methods may be best suited. Electromyographic procedures may be the technique of choice when anatomical specificity is important. Ultrasonic transduction systems have been used and offer an advantage of not requiring attachment of a sensor to the patient, thus preserving the "naturalistic" environment in which to assess dyskinesia.

There are at least two limitations of the present device. First, the device as used in the present configuration, cannot be used to assess perioral dyskinesia. We have reported previously on the efficacy of a similar transducer which may be used to quantify perioral dyskinesia (Caligiuri and Lohr 1989). However, widespread use of this transducer was found not be feasible in a large scale multicenter study because of the added costs necessary to maintain a supply of sterile transducers for each center.

Second, measuring the control of sustained isometric force does not provide information on movement counts which relate more to the clinical description of TD. The sustained force procedure was designed to provide information not readily available to the observer in order to enhance the clinical assessment of TD. Furthermore, the available literature characterizing dyskinetic movements (e.g. frequency of abnormal movement, maximum amplitude of the movement, duration of the movement) does not reveal which of the attributes is best suited as an index of the movement disorder, nor do we know which measure is better suited for a particular purpose (e.g. detection, time course, treatment).

Acknowledgements This study was supported by the Department of Veteran Affairs Cooperative Studies Program. The authors acknowledge the participation of the following personnel (participating investigators; physician participating investigators; research assistants) at the Veterans Affairs Medical Centers (VAMCs): Augusta – Sridhar Gowda, MD; Sukdeb Mukherjee, MD (deceased); Beth Williams. New York – Erica Duncan, MD; Gita Vaid, MD; Marion Schwartz. Northport – Robert Hitzemann, PhD; Ede Frecska, MD; Paul Manicone. Philadelphia – Stanley Caroff, MD; E. Cabrina Campbell, MD; Joan Havey. Portland – Daniel Casey, MD; William Hoffman, MD, and Thomas Hansen, MD; Todd Brundage, Michael Hix (deceased) and Michelle Hanna. San Diego – James Lohr, MD; no PPI; John Browning and Rebecca Vaughan. Tucson – Philip Kanof, MD; no PPI; Robin Fain. West Haven – John Krystal, MD; D. Cyril D'Souza, MD; Donna Damon. West Los Angeles/Brentwood – Steven Marder, MD; Donna Ames, MD; Doreen Ross and Julie Doorack.

Publications resulting from VA Cooperative Study #394 are under the control of the study's Executive Committee: Lenard Adler, MD, and John Rotrosen, MD, Study Co-Chairmen, New York VAMC; Robert Hitzemann, PhD, Participating Investigator, Northport VAMC; James Lohr, MD, Participating Investigator, San Diego VAMC; Robert Edson, MA, Biostatistician, VA Palo Alto Health Care System; Clair Haakenson, RPh, MS (past), and Michael Miller, RPh, MS, Clinical Research Pharmacist, Albuquerque VAMC.

References

Albanese MA, Clarke WR, Adams HP, Woolson RF (1994) Ensuring reliability of outcome measures in multicenter clinical trials of treatments for acute ischemic stroke. The program developed for the Trial of Org 10172 in Acute Stroke Treatment (TOAST). Stroke 25:1746–1751

Bartzokis G, Wirshing WC, Hill MA, et al (1989) Comparison of electromechanical measures and observer ratings of tardive dyskinesia. Psychiatr Res 27:193–198

Bergen JA, Griffiths DA, Rey JM, Beumont PJV (1984) Tardive dyskinesia: fluctuating patient or fluctuating rater. Br J Psychiatry 144:498–502

Bergen JA, Carter NB, Craig J, MacFarlane D, Smith EF, Beumont PJV (1988) AIMS ratings – repeatability. Br J Psychiatry 152:670–673

Caligiuri MP (1997) Instrumental measurements of tardive dyskinesia. In: Yassa R, Nair NPV, Jeste DV (eds) Neuroleptic-induced movement disorders. Cambridge University Press, Cambridge, pp 241–258

Caligiuri MP, Lohr JB (1989) A potential mechanism underlying the voluntary suppression of tardive dyskinesia. J Psychiatr Res 23:257–266

Caligiuri MP, Lohr JB (1990) Fine force instability: a quantitative measure of hand dyskinesia. J Neuropsychiatr Clin Neurosci 2:395–398

Caligiuri MP, Lohr JB, Vaughan RM, McAdams LA (1995) Fluctuation of tardive dyskinesia. Biol Psychiatry 38:336–339

Defazio G, Lepore V, Abbruzzese G et al. (1994) Reliability among neurologists in the severity of assessment of blepharospasm and oromandibular dystonia: a multicenter study. Move Disord 9:616–612

Del Boca FK, Babor TF, McRee B (1994) Reliability enhancement and estimation in multisite clinical trials. J Stud Alcohol [Suppl, Dec 12]:130–136

Edson R, Lavori P, Tracy K, Adler LA, Rotrosen J, and the Veterans Affairs Cooperative Study #394 Study Group (1997) Interrater reliability issues in multi-center trials, part II: statistical procedures used in department of Veterans Affairs Cooperative Study #394. Psychopharmacol Bull 33:59–67

Efron B (1987) Better bootstrap confidence intervals. J Am Statist Assoc 82:171–185

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman and Hall, New York, London

Fleiss JL (1986) The design and analysis of clinical experiments. Wiley, New York

Gardos G, Cole JO, LaBrie R (1977) The assessment of tardive dyskinesia. Arch Gen Psychiatry 34:1206–1212

Gataz WF, Buchel C (1993) Assessment of tardive dyskinesia by means of digital image processing. Psychopharmacology 111:278–284

Henderson WG, Moritz TE, Shroyer AL, et al (1995) An analysis of interobserver reliability and representativeness of data from the Veterans Affairs Cooperative Study on Processes, Structures, and Outcomes in Cardiac Surgery. Medical Care 33:OS86–101

Isaac S, Michael WB (1990) Handbook in research and evaluation, 2nd edn. Edits, San Diego, Calif.

Kane JM, Jeste DV, Barnes TRE, et al (1992) Tardive dyskinesia: a task force Report of the American Psychiatric Association. American Psychiatric Association, Washington

Kim YS, Nibbelink DW, Overall JE (1994) Factor structure and reliability of the Alzheimer's Disease Assessment Scale in a multicenter trial with linopirdine. J Geriatr Psychiatr Neurol 7: 74–83

Kramer MS (1988) Clinical epidemiology and biostatistics. Springer, Berlin

Lane RD, Glazer WM, Hansen TE, et al (1985) Assessment of tardive dyskinesia using the Abnormal Involuntary Movement Scale. J Nerv Ment Dis 173:353–357

Lohr JB, Caligiuri MP (1991) Quantitative instrumental assessment of tardive dyskinesia: a review. Neuropsychopharmacology 6:231–239

Meinert CL (1986) Clinical trials: design, conduct, and analysis. Oxford University Press, New York

Mirra SS, Gearing M, McKeel DW, et al (1994) Interlaboratory comparison of neuropathology assessments in Alzheimer's disease: a study of the Consortium to Establish a Registry for Alzheimer's Disease (CERAD). J Neuropathol Exp Neurol 53: 303–315

Resek G, Haines J, Sainsbury P (1981) An ultrasonic technique for the measurement of tardive dyskinesia. Br J Psychiatry 138:474–478

Richardson MA, Craig JG, Branchey MH (1982) Intra-patient variability in the measurement of tardive dyskinesia. Psychopharmacology 76:269–272

S-Plus for Windows Reference Manual, Volume 2, Version 3.1 (1993) Statistical Sciences, Seattle, Wash.

Simpson GM, Lee JH, Zoubok B, et al (1979) A rating scale for tardive dyskinesia. Psychopharmacology 64:171–179

Smith JM, Kucharski LT, Oswald WT, Waterman LJ (1979) A systematic investigation of tardive dyskinesia in inpatients. Am J Psychiatry 136:918–922

Sweet RA, DeSensi EG, Zubenko GS (1993) Reliability and applicability of movement disorder scales in the elderly. J Neuropsychiatr Clin Neurosci 5:56–60

Trzepac PT, Webb M (1987) The choreometer: an objective test of chorea during voluntary movements. Biol Psychiatry 22: 771–776