J. chron. Dis. 1965, Vol. 18, pp. 303–311. Pergamon Press Ltd. Printed in Great Britain

QUANTITATIVE CLINICAL NEUROLOGICAL TESTING—II*

SOME STATISTICAL CONSIDERATIONS OF A BATTERY OF TESTS

J. W. KUZMA, Ph.D.†, W. W. TOURTELLOTTE, M.D., Ph.D. and R. D. REMINGTON, Ph.D.

The Department of Biostatistics and the Department of Neurology, University of Michigan, Ann Arbor, Michigan

(*Received* 10 *April* 1964)

QUANTITATIVE clinical evaluation of neurological function, e.g., strength, coordination, etc. of humans has so far received less attention than it deserves. This is understandable since the neurological examination is quite complex.

It is no real problem for a neurologist to evaluate the functional capacity of a multiple sclerosis patient and determine if the function tested (e.g., coordination) is supernormal, normal, just subnormal or abnormal (mild, moderate, severe, or no function). These judgments are made on the basis of clinical experience, which is adequate for many purposes. On the other hand, it is most difficult to determine on a subsequent examination of the same patient whether his function is better, worse, or the same; this is frequently true for diseases like multiple sclerosis or cerebrovascular disease even if there is a change which may be recognized by the patient.

Continued advances in many phases of medicine have depended directly upon substituting experimentally verified, objective and quantitative procedures for the classical subjective methods of evaluating biological and neurological phenomena.

To remedy the situation of the lack of objective tests, a number of attempts at a *quantitative* evaluation of neurological disability have been made in recent years. One has been proposed by UNGLEY [1], another by ALEXANDER [2] and a relatively recent one by KURTZKE [3]. These systems permit classification or ranking according to the severity of disability. One objection to such methods is that the ranking is still done in a subjective manner.

There may be a partial solution to this problem of objective assessment. The conventional testing procedures which require clinical judgment for evaluation might be replaced by methods which lend themselves to more objective measure-

*Supported in part by a grant from The Upjohn Company, Kalamazoo, Michigan and The Kenneth H. Campbell Foundation for Neurological Research, Grand Rapids, Michigan.

†USPHS fellow in biostatistics; this paper is a portion of a thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biostatistics, University of Michigan; Present address: UCLA Health Sciences Computing Facility, Los Angeles, California.

Address requests for reprints to: Jan W. Kuzma, Department of Preventive Medicine, University of California, Los Angeles, California, and W. W. Tourtellottee, Department of Neurology, University of Michigan, Ann Arbor, Michigan.

ments so that the data can be analyzed statistically. The lack of more objective methods of evaluating neurological function prompted one of us (W. W. Tourtellotte) to devise over the last several years a battery of tests which measure some of the various types of neurological functions [4]. For example, some of the tests are defined as follows: the strokes per minute on alternate keys of a keyboard counter were used as a measure of a type of coordination; strokes per minute recorded on a single counter were used as a measure of a type of a speed of hand and foot; amplitude of vibrations at 120 cyc/sec discriminated by the finger tips and toes were used as a measure of a type of sensation; the pounds recorded as an instantaneous exertion were used as a type of measure of strength; the pounds recorded per instantaneous exertion on several consecutive trials were used as a measure of a type of a fatigue. The 60 neurological tests (counting each trial separately) included in the battery are given in Table 1.

TABLE 1. BATTERY OF TESTS*

I. Visual acuity

II. Functions of the upper extremity
- (a) Hand speed (5, 10, 20, and 30 sec)
- (b) Hand coordination† (5, 10, 20, and 30 sec)
- (c) Strength (grip, wrist, and deltoid) (2 trials each)
- (d) Fatigue—hand speed (4 trials); grip strength (5 trials)
- (e) Sensation—finger vibration and finger two-point discrimination (3 trials)

III. Functions of the lower extremity
- (a) Station stability—eyes open and eyes closed
- (b) Foot speed (5, 10, 20, and 30 sec)
- (c) Strength—foot and hip flexor (2 trials each)
- (d) Fatigue—foot speed (4 trials); hip flexor strength (5 trials)
- (e) Vibratory sense (toe) (3 trials)

IV. Deep tendon reflexes—jaw, radial, biceps, finger, knee, and ankle

V. Other tests
- (a) Serial sevens
- (b) Bladder urgency

*For a more detailed description of these tests and how they were selected see TOURTELLOTTE *et al.* [4].

†The unadjusted score for coordination consists of the score for the white and red keys. To measure coordination, a subject is asked to strike all the white keys which alternate with red ones. Frequently a subject will strike a red key by mistake. In order to get a score which will reflect coordination more correctly, an adjusted score is obtained by subtracting from the score recorded for the white keys, twice the score recorded for the red keys.

The present study was designed to investigate the reproducibility of the neurological tests under various conditions, to evaluate a learning effect associated with repeated administration of the battery of tests, and to study a fatigue effect associated with repeated administration of the same neurological test.

METHODS

Reproducibility experiment

To provide information as to whether different observers can obtain comparable results, whether the level of neurological function varies during various periods of the day, and whether the level of neurological function varies from day to day in

repeated administration of the battery neurological test variables, an experiment was designed utilizing a 4×4 Graeco-Latin square design with two observations per cell. An example of such a design is given in the appendix. Eight right-handed medical students from the University of Michigan were used in this experiment; a student qualified to enter the study only if his health status was good and free from any history of neurological disorders. The average age was 24 years with a range of 23–25 years; average weight was 182 lb with a range of 165–192 lb. Two of the examiners were neurologists and two were physical therapists. The senior neurologist trained the other three examiners in the test procedures by actually demonstrating the battery of tests to each examiner at one visit.

The neurological tests were administered to only the right-hand side of the body to minimize possible examiner fatigue. Only two students were examined at one sitting by the same examiner. The same set of instruments were used by each examiner. The examinations were performed on four consecutive days during four periods of the day. The four periods were: 9:00–10:45 a.m., 1:00–2:45 p.m., 4:00–5:45 p.m., and 8:00–9:45 p.m. During each period on each day an observer examined two students. During the course of the four days each observer examined each student once and each student was examined once each day but at a different time period on successive days. After a week of rest the entire experiment was repeated using the same observers and the same subjects.

The Graeco-Latin square design is appropriate for this experiment, since it permits repeated use of the same subject and does not require a large number of subjects [5].

New recording sheets were used for each examination to prevent the observer from being influenced by a previously recorded score. The subjects were not permitted to see their score in order to prevent them from trying to beat that score on the next trial.

The following hypotheses were tested:

1. There is no difference between the means of the four observers.
2. There is no difference between the means of the first and second week.
3. There is no difference between the means of the four days in each week.
4. There is no difference between the means of the four periods of the day.
5. There is no difference between the means of the eight students.

A retesting experiment was performed on 27 (of a total of 60) neurological tests which showed significantly different observer means at the 5 per cent level. The experiment was carried out in two parts after the four observers were retrained and the details of administration of the neurological tests were further refined. The first part consisted of 17 (out of 27) neurological tests (tendon reflexes, vibration sense, and 5 sec and 10 sec speed and coordination tests) which could be performed at one sitting without appreciably fatiguing the subjects. A randomized block design using 12 students was utilized for the first part. A new sample of 8 students was used for the second part of the experiment. A Graeco-Latin square design similar to the one of the reproducibility experiment was used.

Learning and fatigue effects

To study learning and fatigue effects an experiment was performed in which a portion of the battery of neurological tests was administered four times to ten

medical students (right and left handed) to both sides of the body. The average age of these students was 31 years with a range of 22–38 years; their average weight was 177 lb with a range of 140–230 lb. Three examinations were administered on consecutive days and the fourth one was administered after three days of rest. Before being admitted to the study the ten subjects were screened using the same medical criteria as the subjects in the above experiment.

Repeated observations on the same subject are correlated. A statistical technique which is appropriate for the analysis of such data was suggested by ELSTON and GRIZZLE [6].

A line was fitted to the observations obtained at the four examinations for each of the neurological tests. The slope of this line was used in testing whether or not the trend (learning) is significant for the four examinations.

This statistical technique was also used in the evaluation of fatigue, the knowledge of which is important to the neurologist. We found that two types of fatigue could be studied from the administration of the neurological tests, one of these being exemplified by grip fatigue which consists of five successive exertions of the hand on a dynamometer. It may be measured by the slope of the line fitted to the observations obtained for the five consecutive trials. The other type of fatigue is associated with successive trials which become more severe due to increased time intervals. Coordination and speed of hand are examples of such tests. This fatigue may be measured by the slope of the line fitted to the number of counts per second in successive trials. The slopes of these fitted lines were used in testing whether or not the trend (fatigue) is significantly different from zero.

It was of interest to investigate whether the mean slope of the two types of fatigue varies significantly from one examination to the next. A two-factor (subjects by examinations) analysis of variance was performed to provide information to answer this question.

RESULTS AND DISCUSSION

There were 27 neurological tests having significantly different observer means at the 5 per cent level, however, this number was reduced to 11 in the re-testing experiment. The results of the re-testing experiment are given in Table 2. The table gives the mean for each observer, the residual mean square from the analysis of variance table, and the F-value. Observers 1 and 2 are the neurologists and observers, 3 and 4 are the two physical therapists.

From Table 2 it may be seen that the differences in observer means for the deep tendon reflex tests are highly significant. The physical therapists were not trained to the point of perfection; this no doubt was a factor in obtaining significant results. A subsequent analysis showed that even the two neurologists who supposedly used the same grading system differed significantly. This may be an indication of what happens when a function is measured without instruments.

Two-point discrimination is just significant at the 5 per cent level. This difference may not be clinically important since the observer means range only from 2.75 to 3.16. It is possible that reading two-point discrimination on a scale of 1 mm intervals may be too fine. Although vibration sense and grip strength gave highly significant F-values for certain trials, it is difficult to imagine why the results were so variable from trial to trial. The same holds true for wrist strength. Wrist strength appears

TABLE 2. LIST OF NEUROLOGICAL TESTS HAVING SIGNIFICANTLY DIFFERENT OBSERVER MEANS IN RE-TESTING EXPERIMENT

Name of neurological test	Observer mean				Residual mean square*	F†
	1	2	3	4		
Deep tendon reflexes‡						
Radial	0.50	1.50	0.50	0.33	0.250	13.70
Biceps	0.33	1.25	0.75	1.42	0.364	7.99
Knee	0.33	1.09	0.08	0.42	0.157	13.92
Finger	0.83	1.58	1.42	0.75	0.254	8.16
Two-point discrimination (mm)	3.16	3.08	3.00	2.75	0.128	3.04
Vibration sense (μ)						
Finger, Trial 1	0.094	0.100	0.127	0.077	0.001	3.85
Finger, Trial 2	0.093	0.087	0.106	0.070	0.001	2.99
Toe, Trial 2	0.540	0.419	0.627	0.477	0.029	3.29
Hand strength (lb)						
Grip, Trial 1	100.7	96.8	104.9	102.1	23.22	3.93
Wrist, Trial 1	51.3	58.2	41.6	42.9	22.91	20.09
Wrist, Trial 2	49.1	55.8	41.6	43.1	12.87	35.80

*The S.E.M. may be obtained by dividing the residual square by the sample size and taking its square root. The sample size is 12 for the first eight tests and 8 for the last three tests.
†The 5 per cent critical value of F with 3 d.f. and 33 d.f. is 2.89 which applies to all but the last three entries. For the last three entries the F-value with 3 d.f. and 15 d.f. is 3.29.
‡Units: 0–4 scale with 2 as normal.

to be the only non-subjective test for which highly significant differences in observer means were found. An explanation of this may be that the examiners are not following the instructions carefully enough in administering this test since a particular examiner may be placing the myometer too far out on the hand thus obtaining a smaller value because of a larger lever effect. This is supported by the consistent results between trials 1 and 2 for any one examiner.

The neurological test variables which showed significantly different means between the four examinations of the first week and the four examinations of the second week at the 5 per cent significance level are given in Table 3. This table gives the means for each week, the residual mean square from the analysis of variance table, and the corresponding F-values.

TABLE 3. LIST OF NEUROLOGICAL TESTS HAVING SIGNIFICANTLY DIFFERENT MEANS BETWEEN WEEKS

List of neurological tests	Means		Residual mean square	F*
	Week 1	Week 2		
Serial sevens (time to finish in sec)	24.63	19.59	40.903	9.90
Bladder: urgency (min)	55.78	94.69	2318.20	10.45
Vibration sense of toe (μ)				
Trial 1	0.64	0.81	0.038	12.43
Trial 2	0.61	0.78	0.049	9.40
Trial 3	0.61	0.76	0.030	11.86
Hand speed for 20 sec (strokes)	138.69	135.69	14.470	9.95
Foot speed for 5 sec (strokes)	28.97	26.97	9.323	6.87
Foot speed for 20 sec	90.16	87.44	23.461	5.04
Deltoid strength; Trial 2 (lb)	60.16	52.09	61.118	6.42
Hip flexion strength				
Trial 1	52.14	57.69	50.714	9.71
Trial 2	50.52	59.21	64.146	19.03

*The 5 per cent critical value of F with 1 d.f. and 30 d.f. is 4.17.

From Table 3 it may be seen that the means of a number of neurological test variables are significantly different for the two weeks. Many of the significant differences, such as speed of hand and foot tests which consist of many repetitions of the same exertion at each examination, could be interpreted as manifestations of a learning effect. The tests consisting of one or two repetitions of one or two actions per examination do not show a learning effect. The serial sevens and bladder emptying tests probably indicate learning effects. The almost two-fold increase in the mean number of minutes of bladder urgency may be explained on the grounds that during the first week the subjects merely respond with a guess, whereas during the second week they give an answer on their bladder function to which their attention has been drawn. Mean vibratory sense of toe is also significantly different for the two weeks. This may be due to the differences in observer technique and instrument error. No significantly different means between periods of day and no significantly different means between days were observed for any of the neurological test variables at the 5 per cent level.

Another experiment was performed using 10 subjects to observe if learning occurred during four examinations (three consecutive exams followed by an additional one three days later). The data from the four examinations were found to fit the linear model (the four examination numbers 1, 2, 3, 4) when the neurological test variables were log transformed. The neurological tests for which a significant trend (regression or learning) was found are given in Table 4. This table gives the slope, its sample variance, the mean square residual from the analysis of variance table, and the F-value on the log scale obtained in time–response curve analysis.

TABLE 4. NEUROLOGICAL TESTS SHOWING SIGNIFICANT REGRESSION ON DATA OF FOUR EXAMINATIONS

Neurological test	Slope	Sample variance of slope	Residual mean square about regression	F*
Coordination of right hand, 10 sec (adjusted)	0.060	0.0006	0.035	6.50
Coordination of left hand, 20 sec (adjusted)	0.031	0.0001	0.016	11.83
Speed of right hand, 30 sec	0.019	0.0001	0.002	6.51
Speed of left foot, 20 sec	0.040	0.0001	0.020	12.46

*The 5 per cent critical value of F with 1 d.f. and 9 d.f. is 5.12.

It may be seen from Table 4 that coordination of hand and speed of hand and foot exhibit a significant learning trend. This learning seems to occur particularly for the 20-sec trials. One may question the relevance of this finding since it is not supported by other similar trials.

This time–response curve analysis was also utilized in studying muscle fatigue associated with repetition of the same trial. It was found that the data fit a linear model when both the neurological test variables and the trial order numbers ($x=1, 2, 3, 4, 5$) have been log transformed. The neurological tests having significant regression (fatigue) at the 5 per cent level are given in Table 5. This table gives the average slope and the estimate of its variance for the ten subjects for each of the four examinations.

TABLE 5. AVERAGE SLOPE AND ITS SAMPLE VARIANCE FOR MUSCLE FATIGUE TESTS FOR TEN MALES

Fatigue test	Examination 1		Examination 2		Examination 3		Examination 4	
	Slope	Sample variance of slope	Slope	Sample variance of slope	Slope	Sample variance of slope	Slope	Sample variance of slope
Grip (lb)								
Right	−0.1917	0.0012	−0.1923	0.0010	−0.1969	0.0004	−0.2377	0.0042
Left	−0.2232	0.0012	−0.2179	0.0006	−0.2171	0.0012	−0.2268	0.0013
Hip (lb)								
Right	−0.0114	0.0429	−0.1975	0.0012	−0.2079	0.0006	−0.1996	0.0006
Left	−0.2434	0.0018	−0.2036	0.0006	−0.2062	0.0024	−0.2458	0.0019

The two-factor analysis of variance used in studying the reproducibility of muscle fatigue in repeated examinations of these tests indicates that the average slopes do not vary significantly at the 5 per cent level.

In studying the fatigue associated with speed of hand and foot in strokes/sec it was found that the data fit a linear model when the neurological test variables had been log transformed. The neurological tests for which significant regression of speed fatigue was found at the 5 per cent level are given in Table 6. This table gives the average slope and the estimate of its variance for the ten subjects for each of the four examinations.

TABLE 6. AVERAGE SLOPE AND ITS SAMPLE VARIANCE FOR SPEED FATIGUE TESTS FOR TEN MALES

Speed test	Examination 1		Examination 2		Examination 3		Examination 4	
	Slope	Sample variance of slope	Slope	Sample variance of slope	Slope	Sample variance of slope	Slope	Sample variance of slope
Hand (strokes/sec)								
Right	−0.0851	0.0001	−0.0803	0.0001	−0.0697	0.0002	−0.0694	0.0001
Left	−0.0477	0.0002	−0.0868	0.0001	−0.0610	0.0001	−0.0843	0.0001
Foot (strokes/sec)								
Right	−0.0413	0.0007	−0.0725	0.0010	−0.0969	0.0007	−0.0686	0.0005
Left	−0.0570	0.0007	−0.0879	0.0002	−0.0793	0.0007	−0.0925	0.0004

The two-factor analysis of variance which was carried out to study the reproducibility of speed fatigue over the four examinations indicates that the average slopes do not vary significantly at the 5 per cent level.

In Table 5 are given the average slopes, which are indications of fatigue associated with five repetitions of the same trial. It was found that the trend of grip fatigue and the trend of hip fatigue for both sides of the body are significantly different from zero and that these estimates of fatigue are stable with repeated administrations of the battery of tests over the period of four days. The same was found for the neurological tests of speed of hand and speed of foot whose values are given in Table 6.

Since these measures of the two types of fatigue are stable, they may be used in the evaluation of neurological disability.

From the experiments on the normal subjects, information was obtained which indicated some of the limitations of the battery and provided clues how some of the tests in the battery should be redefined so that they would be more appropriate for testing multiple sclerosis patients.

It is the opinion of the authors that such a battery of neurological tests as this one, which would provide more objective means of ascertaining neurological function than the conventional neurological examination, should have considerable merit when used in a therapeutic trial. This is especially true in a collaborative study where patients in different hospitals tend to be examined using methods which may not be uniform throughout the different hospitals.

In a publication by TOURTELLOTTE *et al.* [4] descriptive statistics of the various neurological test variables are given for the normal subjects used in this study. A more detailed statistical consideration of the battery of tests may be found in [7].

SUMMARY

A battery of clinical neurological tests was evaluated statistically. In one experiment eight medical students were studied to obtain information on the reproducibility of four observers, eight repeated examinations, and four time periods of day. To obtain learning and fatigue effects another experiment using ten medical students was carried out. The results from this study for these particular groups indicate:

1. That different examiners, such as neurologists and physical therapists, may be trained to obtain comparable results using the quantitative tests of the battery of clinical neurological tests.
2. That the level of neurological function obtained using these neurological tests does not vary significantly during the four stated periods of the day.
3. That the level of neurological function does not differ significantly when the battery of neurological tests was administered on four consecutive days.
4. It is doubtful that a significant learning trend exists when these neurological tests are administered on four consecutive days.
5. That the weekly means of most of the neurological tests which involve many repetitions of the same action at each examination are significantly larger for the second week.
6. The neurological tests of five repeated trials, e.g., gap strength, and the tests of four reduced trials, e.g., speed of hand, may be used as useful measures of two types of fatigue.

REFERENCES

1. UNGLEY, C. C.: Subacute combined degeneration of the cord, *Brain* **72,** 382, 1949.
2. ALEXANDER, L.: New concept of critical steps in chronic debilitating neurologic disease in evaluation of therapeutic response, *Arch. Neurol. Psychiat.* **66,** 253, 1951.
3. KURTZKE, J. E.: On the evaluation of disability in multiple sclerosis, *Neurology* **11,** 686, 1961.
4. TOURTELLOTTE, W. W., HAERER, A. F. and KUZMA, J. W.: Quantitative clinical neurologic testing. I. A battery of tests designed to evaluate in part the neurological function of patients with multiple sclerosis and its use in a therapeutic trial, *N.Y. Acad. Sci.* (In press).
5. COX, D. R.: *Planning of Experiments.* Wiley, New York, 1958.
6. ELSTON, R. C. and GRIZZLE, J. E.: Estimation of the response curves and their confidence bands, *Biometrics* **18,** 148, 1962.
7. KUZMA, J. W.: *A Statistical Study of Various Aspects of a Battery of Clinical Neurologic Tests.* Unpublished doctoral dissertation, University of Michigan, 1963.

APPENDIX

A randomly selected Graeco-Latin square with two observations per cell (a different pair of subjects in each cell) used in the reproducibility experiment.

	1st day	2nd day	3rd day	4th day
Period I	Examiner 1 Subjects 4 & 7	Examiner 2 Subjects 2 & 6	Examiner 4 Subjects 8 & 5	Examiner 3 Subjects 3 & 1
Period II	Examiner 4 Subjects 3 & 6	Examiner 3 Subjects 8 & 7	Examiner 1 Subjects 2 & 1	Examiner 2 Subjects 4 & 5
Period III	Examiner 2 Subjects 8 & 1	Examiner 1 Subjects 3 & 5	Examiner 3 Subjects 4 & 6	Examiner 4 Subjects 2 & 7
Period IV	Examiner 3 Subjects 2 & 5	Examiner 4 Subjects 4 & 1	Examiner 2 Subjects 3 & 7	Examiner 1 Subjects 8 & 6