J. chron. Dis. 1969, Vol. 21, pp. 803–814. Pergamon Press. Printed in Great Britain

AN ASSESSMENT OF THE RELIABILITY OF THREE METHODS USED IN EVALUATING THE STATUS OF MULTIPLE SCLEROSIS PATIENTS*

JAN W. KUZMA,† Ph.D., N. S. NAMEROW†, M.D., W. W. TOURTELLOTTE,‡ M.D., W. A. SIBLEY,¶ M.D., J. F. KURTZKE,§ M.D., A. S. ROSE,† M.D. and W. J. DIXON,† Ph.D.

(*Received* 1 *August* 1968; *in revised form* 3 *October* 1968)

INTRODUCTION

IN THE past few years, increasing effort has been given to the development of quantitative methods for the evaluation of neurologic function so that more objective methods could take the place of subjective methods in the evaluation of therapeutic response. Attempts to estimate the extent and degree of deficits in neurologic function by numerical methods have raised questions about the reliability of such scores. Do they accurately reflect degrees of neurologic dysfunction? More specifically, can such scores be depended upon to portray accurately the changes in degree of a patient's dysfunction from 1 examination to another? Do several examiners, scoring dysfunction for the *same* patient, obtain the *same* score? It may also be asked whether several examiners will observe the same increment of change in a numerical score, representing a change in function between successive examinations of the same patient. These questions are particularly pertinent in collaborative studies where appropriate patient evalution is frequently based on numerical scores and changes in scores as obtained at several investigating centers.

The issues of accuracy and repeatability were particularly relevant to a recent cooperative study evaluating ACTH therapy in Multiple Sclerosis [1]. Four different methods of evaluating patient dysfunction were used, and the separate items of each method were recorded as numerical scores or were changed to numerical scores when the items were recorded as "slight", "moderate", or "severe", etc. Since patients were evaluated by several investigators who submitted data to one statistical center, it was considered essential that a test giving an indication of the reliability of the clinical evaluation methods be performed prior to the analysis of the data.

This report gives the results of the tests of the reliability of 3 of the 4 evaluation methods and presents a statistical design that is considered efficient relative to other

*This study is supported by a grant from the NINDB. Computer assistance is obtained from the Health Sciences Computing Facility, UCLA School of Medicine, sponsored by NIH, grant FR-3.
†Center for Health Sciences, UCLA.
‡Department of Neurology, University of Michigan.
¶Department of Neurology, University of Arizona.
§Veterans Administration Hospital, Washington, D.C.

designs requiring large numbers of cases, and illustrates the use of the design in a reliability study.

METHODS AND MATERIAL

Three of the 4 methods of evaluating dysfunction were used in this experiment. One of these methods was a standard neurologic examination, in which deviations from normal were arbitrarily quantified; the second was the numerical scoring system for each functional system developed by KURTZKE [2]; and the third was a "7-day symptom" scoring method that measured both duration and severity of various individual symptoms over a week's period. The fourth method, a battery of quantitative neurologic tests, was described earlier [3]. The data forms of the 3 methods are given in Appendix 1. For a more detailed account of these methods, see the ACTH-MULTIPLE SCLEROSIS STUDY PROTOCOL [1].

The design of the relability experiment is an incomplete Latin-square, as described by FEDERER [4]. This design, whose layout is shown in Fig. 1, permitted tests of the following hypotheses:

Morning sessions					
	Patient number				
	1	2	3	4	5
Examiner					
one	I	II		III	
two	II	I	III		
three	III		I		II
four			II	I	III
five		III		II	I
Afternoon sessions					
	Patient number				
	6	7	8	9	10
Examiner					
one	I		III		II
two	II	I		III	
three		II	I		III
four	III		II	I	
five		III		II	I

FIG. 1. Roman numerals indicate the sequence of examinations for the multiple sclerosis uniformity study—by patient number and examiner.

(1) There is no difference between the means of the examiners, i.e., on the average, examiners obtain uniform scores on the same patient.

(2) There is no difference between the means of the order of the examinations, i.e., the patient's dysfunction does not change over the period during which the three examinations are given.

(3) There are no differences between the examiner means of the increment change (second-first trial), i.e., changes in dysfunctions, as indicated by numerical scores, are observed uniformly by different examiners.

This design appears appropriate in fitting the following restrictions of the experiments:

(*a*) Patients were not to be examined by more than 3 examiners in a session, to minimize the possibility of patient fatigue.

(*b*) Only 5 examiners were to be used in the experiment.

(*c*) Each examiner was to evaluate an equal number of patients.

The experiment consisted of 4 parts—2 trials, 6 days apart, each trial including a morning and an afternoon session. Five examiners and 5 patients took part in the morning session of the first trial. Each examiner evaluated 3 patients in sequence, and each patient was examined 3 times. Five different patients were seen in the afternoon session, with the same examiners and a similar pattern of testing. Six days later the second trial took place with the same examiners, the same schedule, and the same patients, except for 1 alternate patient in the morning session.

Five neurologists from 4 institutions participated in the study—a senior resident in neurology and 2 junior and 2 senior clinicians, all experienced in the specific methods being used. For this experiment, each of the patients stayed in his assigned room and was visited in sequence by 3 examiners. The time needed to perform the tests was about 30 min, followed by rest periods for patients of 15–30 min. The completed data forms were returned for analysis at the end of each examination, and it was requested of the investigators that they not discuss any results until the end of the study. When an investigator made his second examination of a given patient 6 days later, he could refer to his own first-trial data forms for the patient. This was done to aid in determining whether or not the examiners were observing *changes* in neurologic function uniformly over the period of 1 week, without having to guess the initial level relative to which the second examination was compared. A 6-day interval between 2 examinations of any one patient by any examiner was chosen because it approximates the 1-week interval between examinations that occurs in the cooperative study. Since inter-examiner differences in the initial level and in the change between the initial and the second trial on a single group of patients were the quantitites to be evaluated, the fact that many of the measurements were ordinal, rather than strictly numerical, did not violate the assumptions underlying the statistical procedures.

All of the patients resided in Southern California. Six were males with ages ranging from 33 to 51 yr and a median age of 38 yr. The range of ages for the 5 females was 26–45 yr, with a median of 33 yr. The Kurtzke ratings on the 11 patients ranged from 2–7. Only 1 of the patients was considered to be dynamically changing at the time of the study.

An extended Latin-square design was the means of analyzing the sets of data that were collected for each of the 4 sessions—morning and afternoon of the first day, and morning and afternoon of the second day. The same design was employed in the analysis of 2 more sets of data obtained as the differences between results of the first trial and results of the second trial. The analysis of variance table utilized in this analysis is shown in Appendix II.

Since a patient's neurological status might change even in such a short time as that between the first and third examinations within any 1 of the sessions, a simultaneous examination by all 5 investigators was considered important, and a special session was held. The special session consisted of 3 examinations, during each of which, 1 of 3 patients was examined by a senior clinician while the other four neurologists observed with the exception of some of the neurologic functions such as reflexes for which each examiner had a chance to perform his own examination.

TABLE 1. NEUROLOGIC TESTS WITH SIGNIFICANT DIFFERENCES ($p < 0.01$) IN ADJUSTED EXAMINER MEANS FOR WEEK 1

Test	Adjusted examiner means ($n=10$)					S.E. of the mean	Groups of examiners with significant differences
	1	2	3	4	5		
Neurologic Exam							
Knee-jerk—R	4.4	3.6	3.3	4.1	2.8	0.225	[1, (3, 5)], [4, 5]
Ankle-jerk—R	3.5	2.5	3.8	3.8	2.9	0.197	[2, (1, 3, 4)]
Ankle-jerk—L	3.4	2.1	3.4	3.8	3.5	0.283	[2, (1, 3, 4, 5)]
Limb spasticity—RU	−0.03	0.5	0.03	−0.03	0.03	0.100	[2, (1, 3, 4, 5)]
Limb spasticity—LU	0.13	0.53	0.0	0.0	0.0	0.114	[2, (3, 4, 5)]
Limb spasticity—LL	0.40	1.50	0.53	0.13	0.47	0.240	[2, (1, 4)]
Sensory Deficit							
Vibratory—RL	1.5	2.3	1.2	2.3	1.5	0.234	[3, (2, 4)]
Superficial—RL	1.4	0.30	0.30	−0.10	0.30	0.202	[1, (2, 3, 4, 5)]
Superficial—LL	1.5	0.10	0.23	0.17	0.17	0.246	[1, (2, 3, 4, 5)]
Superficial—Trunk	1.0	0.10	0.23	0.03	0.10	0.187	[1, (2, 4, 5)]
Neurologic Status Evaluation							
Bowel and bladder	2.6	1.8	2.6	1.5	1.6	0.187	[1, (4, 5)], [3, (4, 5)]
Disability status scale	5.0	3.9	5.1	3.6	3.8	0.427	[3, 4]
7-day Symptom Scoring							
Numbness—RL	8.9	4.4	5.7	4.3	4.7	0.830	[1, (2, 4, 5)]
Bowel and bladder	13.6	12.2	16.9	10.4	10.4	1.109	[3, (4, 5)]

None of these three patients had taken part in any other session; 2 were females with Kurtzke scores of 5 and 6, and one was a male who scored 2. Although the conditions of this experiment differed markedly from those at the doctors' hospitals, the experiment provided an opportunity to see how uniformly *all* 5 of the examiners graded the various neurologic dysfunctions. Appendix III shows the analysis of variance table for the randomized-blocks design used for the special examinations.

RESULTS

Significant differences ($p<0.01$) between examiner means for the *first week's examination* were observed for 14 of the 87 items evaluated. Ten of these items were from the neurologic examination, 2 from the neurologic status evaluation, and 2 from the 7-day symptom scoring method. Since each examiner observed a slightly different group of patients from that observed by the other examiners, an adjustment was made to the examiner means to permit valid comparisons among the 5 examiners. The adjusted examiner means and the pairs of examiners for which significant difference were found in a multiple range test, as described by STEEL and TORRIE [5], are shown in Table 1. These results indicate that most of the differences occurred between examiner 1 or examiner 2 and the remaining examiners. Specifically, examiner 1 differed from other examiners on the right knee-jerk reflex, superficial sense deficits of legs and trunk, numbness of the right leg, and bowel and bladder function, while examiner 2 differed from the others on the left ankle-jerk reflex and spasticity of 3 limbs. Other tests of the limbs did not show any such discrepancies in reflexes, tone, or sensation. This observation—that each difference is primarily between 1 examiner and the others—implies that the others were able to obtain reasonably comparable scores.

For the *second week's examination*, significant differences ($p<0.01$) between examiner means were observed for only 10 of the 87 items. Nine of these were from the neurologic examination and 1 was from the 7-day symptom scoring method. The adjusted examiner means and the pairs of examiners for which significant differences were observed in a multiple range test are shown in Table 2. These results indicate that the significant differences during the second week were primarily between pairs of observers rather than between any 1 and the other 4. Examiner 2 differs from the others on both of the triceps reflexes, and he has the highest score for spasticity of the left leg. Examiner 1 differs from the other examiners with respect to several measures of sensory deficit. Significant differences during the second week, though still attributable to the same 2 examiners as in the first week, were less frequent and were mostly attributable to different items.

A comparison of Table 2 with Table 1 shows that significant differences were found for both trials in only 5 specific tests—the superficial sense deficits for the trunk and both legs, spasticity for the left leg, and "numbness" for the right leg. Note that the other body parts did not show differences for these items.

To indicate the precision of the tests among the adjusted examiner means, standard errors of the means are given in each of the Tables. Note that these standard errors are for the discrepant values; those for items without significant differences were much smaller. The finding of the few statistically significant differences takes on limited importance in view of the additional information that only 5 of the tests gave significant differences at both the first and second trials, that these differences

TABLE 2. NEUROLOGIC TESTS WITH SIGNIFICANT DIFFERENCES ($p<0.01$) IN ADJUSTED EXAMINER MEANS FOR WEEK 2

Test	Adjusted examiner means ($n=10$)					s.e. of the mean	Groups of examiners with significant differences
	1	2	3	4	5		
Neurologic Exam							
Triceps—R	2.5	1.9	2.8	3.0	2.7	0.218	[2, 4]
Triceps—L	2.6	2.0	2.5	3.0	2.8	0.147	[2, 4]
Limb spasticity—LL	0.37	1.2	0.63	−0.03	0.63	0.202	[2, 4]
Sensory Deficit:							
Vibratory—RU	−0.13	0.80	1.07	0.80	1.47	0.260	[1, 5]
Position—RL	1.1	0.30	0.43	0.17	0.50	0.144	[1, 4]
Position—LL	1.07	0.20	0.13	0.13	0.13	0.134	[1, (2, 3, 4, 5)]
Superficial—RL	1.23	0.23	0.43	−0.03	0.30	0.187	[1, 4]
Superficial—LL	1.27	−0.07	0.20	0.07	0.20	0.158	[1, (2, 3, 4, 5)]
Superficial—Trunk	0.80	0.07	0.13	−0.07	0.07	0.126	[1, (2, 4, 5)]
7-day Symptom Scoring							
Numbness—RL	9.8	4.0	5.4	4.6	6.0	0.881	[1, (2, 4)]

were primarily due to 1 or 2 examiners, and that the differences were observed for only a few individual tests rather than for related groups of functions. Furthermore, the precision of the methods, as indicated by the small values obtained for the standard errors of the means, was high.

For testing the *changes between the first and second trial for the different examiners*, the results of the first week's examinations were subtracted from those of the second and an adjusted average ($n=10$) for the 5 examiners obtained. The only item for which a significant difference ($p<0.01$) in the average examiner differences occurred was the right ankle-jerk reflex, a difference due to examiners 2 and 4. The results of the remaining 86 tests were essentially uniform. Such uniformity of the differences (second-first trial) among the examiners has particular importance since the differences (second-first trial) rather than differences in the level of the functions are used to evaluate the efficacy of treatment, notably in the ACTH study.

From the analysis of examinations with respect to *order of examination*—the order in which each patient's sequence of 3 took place—only 1 significant difference ($p<0.01$) was found among the 87 items, that for the bowel and bladder function. Interrogation of the patient was the means of obtaining the result of this test. For this function, the average score for the second examination was significantly higher than those for the first and third tests.

The analysis of the results of the 3 *special examinations*, in which all 5 examiners participated simultaneously, showed that there were no significant differences ($p<0.01$) in examiner means for any of the neurologic functions tested.

The observation of very few statistically significant differences in this entire study takes on added importance in view of the facts that only 5 of the items gave any significant differences at both the first week's and the second week's trials, that these differences were primarily due to 1 or 2 examiners, that only 1 item out of 87 showed a significant difference in the degree of change between the 2 examinations, and that there were no significant examiner differences for the special examinations.

DISCUSSION

The assessment of treatment results in multiple sclerosis presents various problems. To be useful in demonstrating statistically significant trends or differences, assessment methods must be shown to be reliable. The results of this study—that the 5 examiners did not differ significantly in 82 of the 87 items included in the 3 scoring methods—indicate that the methods evaluated can be used in obtaining appropriate data on neurological dysfunction and on changes of function in multiple sclerosis patients. Some of the significant differences that were observed may have been due to chance as they appeared sporadically—such as reflexes on one side but not the other—and were not present on repeated testing. Significant differences were observed for only 1 or 2 of the 5 examiners on even these deviant items. Most importantly, although differences in the level of a neurologic function occurred on several items, differences between examiners did not occur when the results of the first week's tests were compared to those of the second week. This indicates that even if different examiners tend to observe several functions at different levels, they do appear to observe changes in neurologic function uniformly. And it is this reliability in the evaluation of increment change that is most pertinent to the conduct of a therapeutic trial.

This study illustrates the use of an efficient statistical design, which can be used

to test whether or not results from various investigators in a cooperative trial can be pooled for valid analysis. The incomplete Latin-square design used in this study required only $2\times5\times3=30$ observations, whereas, a comparable 3-factor analysis of variance design would have required $10\times5\times3=150$ observations. A test of the hypothesis that there is no difference in the order of examinations gave support to the assumption that a sequence of 3 examinations would not significantly alter a patient's neurologic functions. The finding of the test was that there were no significant differences in the sequences of the 3 examinations except for one of the 87 items measured, and even here, it was the second exam that differed from the other 2. Since each patient was examined only 3 times instead of 5—a precaution taken to minimize the effects of fatigue on repeated assessment of the neurologic function of a single patient—the 5×3 Latin-square design is classed as incomplete.

Utilizing such a design it is possible to obtain information on the uniformity of pooled data and to make statements regarding the reliability of the data for the various neurologic functions. Specifically, for this ACTH study it will be possible to indicate the variables for which statistical analysis should be meaningful. When such precautions are included in the design and analysis of data, unwarranted inferences will be minimized and progress in the evaluation of such chronic diseases as multiple sclerosis may continue.

SUMMARY

The reliability of three different evaluation methods used in a cooperative clinical trial of the efficacy of ACTH in multiple sclerosis patients was evaluated in a uniformity study that used an efficient statistical design requiring only 10 patients and 5 examiners. The methods were the standard neurologic examination, a scoring system for functional grades and disability status, and a 7-day symptom score. Each patient was examined only 3 times at the beginning of the study and 3 more times 6 days later. No significant differences among the 5 examiners were observed on 82 of the 87 items used to measure neurologic function. With the exception of 1 variable, there were no significant differences among the average values of the sequence of the 3 examinations, nor among the average increments of change in the numerical scores between the first and second trials.

In an additional examination in which all 5 examiners simultaneously evaluated 3 patients 1 at a time, it was found that the 5 examiners observed uniformly in all of the neurologic tests.

The results of this study indicate that, by and large, the three evaluation methods appear to be reliable in the evaluation of neurologic status when used in a cooperative clinical trial where several investigators contribute data. Furthermore, investigations of reliability in cooperative studies can be performed with the use of efficient statistical designs such as the incomplete Latin-square design.

REFERENCES

1. ROSE, A. S. *et al.:* Cooperative study on the evaluation of therapy in multiple sclerosis; ACTH vs. placebo in acute exacerbations. Preliminary Report. *Neurology* **18,** (2), June 1968.
2. KURTZKE, J. F.: Further notes on disability evaluation in multiple sclerosis with scale modifications. *Neurology* **15,** 654, 1965.
3. KUZMA, J. W. *et al.:* Quantitative clinical neurologic testing. II. Some statistical considerations of a battery of tests. *J. chron. Dis.* **18,** 303, 1965.

4. FEDERER, W. T.: *Experimental Design. Theory and Application.* McMillan, New York, 1955.
5. STEEL, R. G. and TORRIE, J. H.: *Principles and Procedures of Statistics.* McGraw-Hill, New York, 1960.

APPENDIX I

A. Neurologic examination

1. Vision: Acuity, R and L (Snellen Chart)
   Field Abnormalities, R and L, Yes or No

2. Reflexes: (0–5+, 2+: average)
   Biceps, R and L
   Triceps, R and L
   Radial, R and L
   Knee-jerk, R and L
   Ankle-jerk, R and L
   Plantar; Extensor, R and L
   Neutral, R and L
   Flexor, R and L
   Snout; present, Yes or No
   Jaw; abnormal, Yes or No
   Abdominals; present, R and L, Yes or No

3. Brain Stem—Signs of Abnormalities:
   (0=none, 1=slight, 2=moderate, 3=severe, 4=total)
   Facial weakness
   Facial sensory loss
   Nystagmus
   EOM and/or gaze impairment
   Dysarthria
   Dysphagia
   Other bulbar signs

4. Limb Weakness:
   (0=none, 1=slight, 2=moderate, 3=severe, 4=total)
   RU, RL, LU, LL

5. Limb Spasticity:
   (0=none, 1=slight, 2=moderate, 3=severe, 4=total)
   RU, RL, LU, LL

6. Limb Coordination Impairment:
   (0=none, 1=slight, 2=moderate, 3=severe, 4=total)
   RU, RL, LU, LL

7. Sensory Deficit:
   (0=none, 1=slight, 2=moderate, 3=severe, 4=total)
   Vibratory; RU, RL, LU, LL
   Position; RU, RL, LU, LL
   Superficial; RU, RL, LU, LL. Trunk

8. Gait Impairment:
   (0=none, 1=slight, 2=moderate, 3=severe, 4=total)

9. Other: (0=none, 1=slight, 2=moderate, 3=severe, 4=total)

10. Mental Status:
    (0=none, 1=slight, 2=moderate, 3=severe, 4=total)
    Mood Abnormalities; Depression
    Elation
    Anxiety
    Euphoria
    Intellect Impairment
    Morbid Ideas

B. Neurologic Status Evaluation

1. Pyramidal functions:
   - 0=normal
   - 1=abnormal signs without disability
   - 2=miminal disability
   - 3=mild or moderate paraparesis or hemiparesis; severe monoparesis
   - 4=marked paraparesis or hemiparesis, moderate quadriparesis; or monoplegia
   - 5=paraplegia, hemiplegia, or marked quadriparesis
   - 6=quadriplegia

2. Cerebellar functions:
   - 0=normal
   - 1=abnormal signs without disability
   - 2=mild ataxia
   - 3=moderate truncal or limb ataxia
   - 4=severe ataxia all limbs
   - 5=unable to perform coordinated movements due to ataxia

3. Brain stem functions:
   - 0=normal
   - 1=signs only
   - 2=moderate nystagmus or other mild disability
   - 3=severe nystagmus, marked extraocular weakness, or moderate disability of other cranial nerves
   - 4=marked dysarthria or other marked disability
   - 5=inability to swallow or speak

4. Sensory functions:
   - 0=normal
   - 1=vibration or figure-writing decrease only 1–2 limbs
   - 2=vibration decrease 3–4 limbs, or position or discrimination decrease 1–2 limbs
   - 3=mild decrease touch, pain; or loss position, vibration 1–2 limbs
   - 4=moderate decrease touch, pain for at least most of 1 limb; severe proprioceptive decrease 3–4 limbs
   - 5=loss of sensation for 1 limb, or moderate decrease touch, pain of most of body
   - 6=analgesia and anesthesia to neck

5. Bowel and Bladder functions:
   - 0=normal
   - 1=mild hesitancy, urgency, or retention
   - 2=moderate hesitancy, urgency, retention, or rare urinary incontinence
   - 3=frequent incontinence
   - 4=in need or almost constant catheterization but with adequate bowel function
   - 5=loss of bowel and bladder function

6. Visual functions:
   - 0=normal
   - 1=scotoma with visual acuity (corrected) better than 20/30
   - 2=worse eye with scotoma with maximal visual acuity (corrected) of 20/59
   - 3=worse eye with large scotoma, or moderate decrease in fields, but with maximal visual acuity (corrected) of 20/60 to 20/99
   - 4=worse eye with marked decrease of fields and maximal visual acuity (corrected) of 20/100 to 20/200; grade 3 plus maximal acuity of better eye 20/60 or less
   - 5=worse eye with maximal visual acuity (corrected) less than 20/200; grade 4 plus maximal acuity of better eye 20/60 or less
   - 6=grade 5 plus maximal visual acuity of better eye 20/60 or less

7. Mental functions
   0=normal
   1=mood alternation only
   2=mild decrease mentation
   3=moderate decrease mentation
   4=marked decrease mentation (chronic brain syndrome, moderate)
   5=dementia; or chronic brain syndrome, severe, incompetent

8. Other functions:
   0=none
   1=any other findings (specify)

9. Disability Status Scale in Multiple Sclerosis:
   0=normal neurologic exam (all grade 0 in functional systems)
   1=no disability, minimal signs
   2=minimal disability
   3=moderate disability though fully ambulatory
   4=relatively severe disability though fully ambulatory and able to be self-sufficient and up and about for some 12 hr a day
   5=disability severe enough to preclude ability to work a full day without special provisions; maximal motor function walking unaided no more than several blocks
   6=assistance required for walking
   7=restricted to wheel chair—able to wheel self and enter and leave chair alone
   8=restricted to bed but with effective use of arms
   9=totally helpless bed patients
   10=death due to multiple sclerosis

C. Seven-day symptom scoring

		Number of days of severity during last 7 days				
		Slight 1 x d +	Moderate 2 x d +	Severe 3 x d*	Is symptom part of current bout ? Yes	No
1. Extremity Symptoms						
Number of days of weakness or incordination or cerebellar ataxia during last 7 days	R. arm					
	L. arm					
	R. leg					
	L. leg					
Number of days of numbness and/or paresthesias during last 7 days	R. arm					
	L. arm					
	R. leg					
	L. leg					
2. Visual Symptoms						
Number of days of impaired vision due to retrobulbar neuritis during the last 7 days	R.eye					
	L.eye					

* d= number of days

Gait ataxia					
Girdle sense					
Bladder and/or bowel					
Diplopia					
Facial numbness					
Facial weakness					
Dysarthria					
Dysphagia					
Vertigo					
Altered Hearing					

APPENDIX II

Analysis of variance Table of extended incomplete Latin-squares

Source		Degrees of freedom	Sum of squares*	Mean squares
Columns	Replicate—a.m.+p.m.	$r-1$ $\begin{cases} 1 \\ r-2=1 \end{cases}$	C $\begin{cases} R \\ O \end{cases}$	R/d.f.
	Order			O/d.f.
	Block—patients (adjusted)	$b-1=9$	B	B/d.f.
	Treatments—examiners (unadjusted)	$t-1=4$	T	T/d.f.
	Error	$(r-1)(b-1)-(t-1)=14$	E	E/d.f.
	Total	$rb-1=29$		

*Entries for this column are defined as follows:

1. Calculate the column totals C_i, the block totals B_j, the treatment totals T_k, the replicate totals R_m, and the grand total G.

2. (*a*) For replicate sum of squares $R=\dfrac{\Sigma R_m^2}{rt}-\dfrac{G^2}{rb}$

   (*b*) For column sum of squares $C=\dfrac{\Sigma C_i^2}{b}-\dfrac{G^2}{rb}$

   (*c*) For order sum of squares $O=C-R$

3. For treatment sum of squares (unadjusted) $T=\dfrac{\Sigma T_k^2}{2r}-\dfrac{G^2}{rb}$

   Compute for each treatment the quantity $Q_j=KT_k-B_{t_k}$, where B_{t_k}=total for all blocks in which treatment t_k appears, and K is the number of times a patient is examined. For treatment sum of squares (adjusted) $T_a=\dfrac{1}{tK\lambda}\Sigma Q_j^2$, where t is the number of treatments, and λ is the number of times that each pair of examiners occurs together.

4. For block sum of squares (unadjusted) $B_u=\dfrac{\Sigma B_j^2}{b}-\dfrac{G^2}{rb}$, and for block sum of equares (adjusted) $B=B_u+T_a-T$.

5. For error sum of squares $E=\sum_{i}^{r}\sum_{j}^{b}X_{ij}^2-(C+B+T)$.

APPENDIX III

Analysis of variance Table of randomized blocks design

Source	Degrees of freedom	*Sum of squares*
Patients	$r-1=2$	$c\sum_{i=1}^{r}(\bar{y}_{i.}-\bar{y})^2$
Examiners	$c-1=4$	$r\sum_{j=1}^{c}(\bar{y}_{.j}-\bar{y})^2$
Error	$(r-1)(c-1)=8$	$\sum_{i=1}^{r}\sum_{j=1}^{c}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y})^2$
Total	$rc-1=14$	$\sum_{i=1}^{r}\sum_{j=1}^{c}(y_{ij}-\bar{y})^2$