INTERNATIONAL JOURNAL OF LEPROSY Volume 63, Number 1
Printed in the U.S.A.

Inter-Observer Variability in the Assessment of Nerve Function in Leprosy Patients in Ethiopia[1]

Christian Lienhardt, Heather Currie, and Jeremy G. Wheeler[2]

Leprosy causes disability and deformity through damage to peripheral nerves. *Mycobacterium leprae* has the unique characteristic of entering peripheral nerves and multiplying within Schwann cells, but the host's response to this invasion is extremely variable: it can be minimal with no functional changes in the nerve, or it may be very extensive, resulting in a severe loss of function, with a risk of severe disability ([14]). In addition, acute episodes of neuritis causing nerve destruction can occur throughout the course of leprosy, mainly related to type 1 or type 2 reactions ([12, 14, 17]).

Clinically, neuritis is characterized by pain and/or tenderness in the nerves, accompanied by a loss of sensory or motor function ([23]), but can sometimes be silent, with no noticeable signs and symptoms ([9]). The relationship between neuritis and nerve damage is complex since there can be neuritis with little or no evidence of nerve damage or vice versa ([14]). In both situations, however, there is a risk of irreversible disability or deformity as a consequence of anesthesia, dryness of the skin and/or muscular paralysis in various combinations ([8]).

In 1985, the World Health Organization (WHO) identified prevention of disability as one of the main objectives of leprosy control and encouraged leprosy control programs (LCPs) to focus on early detection and treatment of nerve damage ([27]). To achieve it, leprosy patients must be monitored regularly and carefully during and after treatment, in order to follow properly their evolution and to detect any new sign of neuritis and/or nerve damage ([4, 22]).

Several tests have been developed to assess and grade motor and sensory functions in leprosy patients. With respect to sensory function, various methods have been developed. The most in use are the Semmes-Weinstein nylon monofilaments ([24]), which assess the sensory response to an increasing range of predetermined forces ([1, 13, 19, 22]) and have been reported to present good repeatability in clinical testing ([2, 3]), and the ballpoint pen ([25]), which tests the presence or absence of sensory response to a single stimulus. Although the latter method is less standardized and sensitive, it is widely used in LCPs, especially in the field, because of its simplicity and low cost. Concerning motor function, Goodwin ([11]) developed in 1968 voluntary motor testing (VMT) for leprosy patients, based on the U.K. Medical Research Council (MRC) scale of strength (Medical Research Council. Aids to the investigation of peripheral nerve injuries. MRC Memo No. 7; London, 1943, HMSO.), which subsequently has been reviewed by several authors ([6, 19, 20, 22]). Several scales have been proposed to grade the motor function; the most frequently used is the MRC 5-point scale, but simpler 3- or 4-point scales have been devised, mainly for field use ([25, 26]).

Because patients are followed over a long period of time in LCPs, the continuous monitoring of nerve function requires the use of reproducible and reliable tests, in order to keep variation between observers to a minimum and to avoid misclassification of the patient's neurological status. This study was designed to estimate the variability between observers when performing nerve function tests in leprosy patients under field conditions.

SUBJECTS AND METHODS

The study took place in the All Africa Leprosy Rehabilitation and Training Centre

[1] Received for publication on 11 January 1994; accepted for publication in revised form on 19 October 1994.

[2] C. Lienhardt, M.D., D.T.M., M.Sc., Communicable Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, London, U.K. and Institut Marchoux, Bamako, Mali. H. Currie, All Africa Leprosy and Rehabilitation Center (ALERT), Addis Ababa, Ethiopia. J. G. Wheeler, M.Sc., EPS/CDEU, London School of Hygiene and Tropical Medicine, London WC1E 7HT, U.K.

Reprint requests to J. G. Wheeler.

(ALERT) in Addis Ababa, Ethiopia, in July 1992 and in January 1993.

The assessment of nerve function included: a) detection of clinical signs of neuritis (NS): pain, tenderness and enlargement, recorded on a 3-point scale for each nerve (Annex 1); b) assessment of sensory function with a set of standard nylon filaments (NF) on a 15-point scale and with a ballpoint pen (BP) on a 3- or 5-point scale (Annex 2); and c) assessment of the motor function with an abridged version of voluntary motor testing (VMT) on a 6- or 4-point scale (Annex 3).

Two series of 50 leprosy patients (100 in total) presenting at three different units in ALERT were recruited for this study. After carefully recording their characteristics, the patients were assigned to two examiners who first assessed the signs of neuritis and then performed the nerve function tests (NF, BP, VMT) on each of them. The order in which a patient saw each examiner was determined at random. Measurements of nerve function were thus taken on the same subject by two different observers on different occasions. In order to avoid influencing the results obtained with one sensory test by those obtained with the other test, and to avoid patient fatigue, a time span of 2 hr was allowed between the NF and BP tests. For each test, a specific score was calculated for each single nerve.

Because inter-observer variability is likely to depend upon the qualification and training of the observers, we investigated it with two different types of leprosy workers with different training and experience: two physio-technicians from the ALERT Rehabilitation Unit who routinely assess nerve function in all hospitalized and disabled patients at ALERT, and two health assistants working in field clinics, who are mainly in charge of patient follow up during and after chemotherapy. The scores used were similar for the two groups of observers, except for the VMT, coded according to the MRC scale for the physio-technicians and according to an abbreviated scale (SRMP) for the health assistants (see Annex 3 and Table 1).

Analysis. Systematic variation in measurement of nerve function may be due to differences between observers, instrument and times of measurements ("occasions"). In this study, we assumed that there was no variation of nerve function within a short time period (i.e., no within-subject variation) and that within-observer variation was randomly distributed. With these assumptions, using a balanced design between the two observers to control for the "occasion" effect (i.e., giving an equal chance for each patient to be tested first by observer A or observer B), and using calibrated instruments and/or standardized methods for the measurements, the observed differences between measurements were considered to reflect the inter-observer variation.

For each test, the measurements obtained on each nerve by each observer were cross-tabulated. Agreement was represented by the percentage of measurements which were identical for the two observers. Percentage disagreements by 1 point also were calculated. In order to evaluate whether the degree of agreement between the two examiners was better or not than that predicted by chance alone, weighted kappa ($w\kappa$) statistics were calculated (Annex 4). A value of $w\kappa = 1$ represents perfect agreement, while $w\kappa = 0$ indicates that the agreement is no better than would arise purely by chance. By convention, values below 0.4 indicate poor agreement beyond chance, values between 0.4 and 0.75 represent fair-to-good agreement beyond chance, and values greater than 0.75, excellent agreement beyond chance ([10]).

Another aspect of observer variation is "systematic difference": disagreement between observers occurring in one direction (e.g., observer A tending to rate higher than observer B). Systematic difference was tested by comparing the average difference in score between the two observers to an expected value of zero, using the Wilcoxon signed rank test ([15]).

Some authors have used aggregated scores for nerve function summing over nerves for each test ([19, 22]). We extended our analysis of inter-observer variability, summing up for each test the results obtained for all nerves tested in each patient, in a global score per test (Table 1). This process allowed the transformation of categorical measurements into a pseudo continuous measurement, thus providing us with a summary of the overall variation between examiners for those tests. For comparison purposes, we plotted, for each test, the dif-

TABLE 1. *Scoring system used in the study.*

a. Score per nerve

Nerve	NF[a]	BP[b]	VMT[c]		NS[f]
			PT[d]	HA[e]	
Ulnar	0–15	0–3	0–5	0–3	0–6
Median	0–15	0–3	0–5	0–3	0–6
Radial	—	—	0–5	0–3	0–6
Common peroneal	—	—	0–5	0–3	0–6
Posterior tibial	0–15	0–5	—	—	0–6

b. Score per test

	NF	BP	VMT	NS
PT	0–90	0–22	0–50	0–60
HA	0–90	0–22	0–30	0–60

[a] NF = Nylon monofilaments.
[b] BP = Ball-point pen.
[c] VMT = Voluntary motor testing.
[d] PT = Physio-technician.
[e] HA = Health assistant.
[f] NS = Signs of neuritis.

ferences between the measurements obtained by the two examiners against their mean, according to the method proposed by Bland and Altman ([5]). If there is no systematic difference between observers, the mean difference in measurements should be zero and the difference in measurements should be unrelated to the mean score (i.e., agreement should not depend on severity of nerve damage). Since the score per test is not a true continuous measure, we only present here the descriptive information provided by the plots, and refrain from using formal statistical tests for continuous data to assess inter-observer variability on the global score per test.

RESULTS

Physio-technicians

Description of data. Fifty patients were examined by the two physio-technicians: 20 (40.0%) females and 30 (60.0%) males. Mean age was 31.8 years old; 28.6 years for females, 34.1 years for males. The distribution of leprosy by type was as follows: TT 1 (2%), BT 16 (32%), BT/BB: 1 (2%), BB 1 (2%), BL 22 (44%), BL/LL 2 (4%), LL 2 (4%), other or unknown 5 (10%). Fifteen patients (30%) were in reaction and/or were experiencing neuritis.

Results per test. A summary of the results obtained by the two physio-technicians for each nerve are given in Table 2. The results are examined for each test successively.

NF test. The percent agreement in the NF test ranged from 32% to 58%. In 18% to 28% of the cases the observers disagreed by 1 point over a total score of 0 to 15. The mean difference between measurements was negative for all nerves, indicating that the second observer was constantly rating higher than the first. The evidence for systematic difference between observers was significant for the median and the posterior-tibial nerves ($p < 0.05$). Weighted κ statistics showed good-to-very-good agreement between examiners better than expected by chance alone ($0.736 < w\kappa < 0.814$), but with wide confidence intervals, consistent with both poor agreement and very good agreement beyond that expected purely by chance.

BP test. The percent agreement in the BP test ranged from 71% to 84%. In 16% to 29% of the cases, the examiners disagreed by 1 point or more on a 3- or 5-point scale. The mean differences in score were all positive and were significantly different from zero for the right ulnar, the right median, and the right and left posterior tibial nerve, suggesting a systematic difference between examiners for those nerves. According to $w\kappa$ values, there was a fair-to-good agreement between observers, better than expected by chance alone. Confidence intervals were wider for the feet than for the hands.

VMT. Except for the facial nerve, the percent agreement with VMT was good for all nerves, ranging from 79% to 98%. Mean differences in score varied in direction but were not significant ($p > 0.1$). Despite the high percent agreement, $w\kappa$ values are extremely variable (very low for the facial nerve, high for the ulnar and median nerves), and the wide confidence intervals include values consistent with both perfect agreement or pure chance agreement [The zero weighted kappa value for the right radial nerve was due to the fact that one observer always gave the same rating, and in this situation weighted kappa could not be determined (see Discussion)].

NS. The clinical signs of neuritis agreement was poor for all nerves (13% to 41%). The mean differences were all positive, and

TABLE 2. *Results of nerve function assessment for all nerves by the two physio-technicians.*

Nerve/test	No.	% Agreement	% Disagreement of 1 point	Mean difference	p[a]	Weighted κ	CI of weighted κ
NF							
r ulnar	50	48	20	-0.34	0.271	0.797	0.41 - 1.17
l ulnar	49	51	18	-0.04	0.607	0.814	0.47 - 1.15
r median	50	58	18	-0.62	0.007	0.807	0.28 - 1.32
l median	50	42	28	-0.74	0.026	0.742	0.22 - 1.26
r post-t	50	32	26	-0.50	0.047	0.775	0.35 - 1.19
l post-t	50	38	20	-0.74	0.013	0.736	0.26 - 1.21
BP							
r ulnar	50	78	16	0.44	0.037	0.720	0.53 - 0.88
l ulnar	50	82	10	0.18	0.156	0.781	0.64 - 0.92
r median	50	84	6	0.30	0.007	0.604	0.37 - 0.84
l median	50	80	10	0.18	0.171	0.627	0.41 - 0.84
r post-t	49	69	20	0.29	0.003	0.793	0.44 - 1.15
l post-t	49	71	16	0.39	0.025	0.746	0.34 - 1.15
VMT							
r facial	48	77	19	-0.04	0.340	0.248	-2.43 - 2.92
l facial	48	73	21	-0.15	0.290	0.187	-2.02 - 2.39
r ulnar	49	84	8	-0.24	0.148	0.741	-0.01 - 1.48
l ulnar	49	79	10	-0.02	0.880	0.752	0.30 - 1.20
r median	49	92	6	-0.06	0.500	0.750	-2.10 - 3.60
l median	49	94	-	-0.18	0.250	0.780	-0.51 - 2.07
r radial	49	96	2	0.12	0.500	0	-1.64 - 1.64
l radial	49	96	2	0.06	1.000	0.656	-3.71 - 5.03
r c per	49	98	-	0.10	1.000	0.535	-5.74 - 6.81
l c per	49	92	2	0.08	0.750	0.477	-3.00 - 3.96
NS							
r ulnar	50	20	26	1.58	0.001	0.219	-[b]
l ulnar	50	14	28	1.78	0.001	0.096	-
r median	50	30	28	1.16	0.001	0.030	-
l median	49	41	26	1.02	0.001	0.137	-
r radial	50	28	20	1.52	0.001	0.112	-
l radial	50	20	38	1.08	0.001	0.072	-
r c per	50	22	18	1.54	0.001	0.015	-
l c per	50	18	26	1.72	0.001	0.017	-
r post-t	47	21	13	1.94	0.001	0.100	-
l post-t	47	13	15	2.09	0.001	0.047	-

[a] Tables 2 and 3 display mean differences together with the results (p value) of the Wilcoxon test, which is based on the median difference. Means rather than medians are displayed as the sign of the mean difference between observers was informative as to the direction of the bias.

[b] Standard error could not be calculated due to negative square root values.

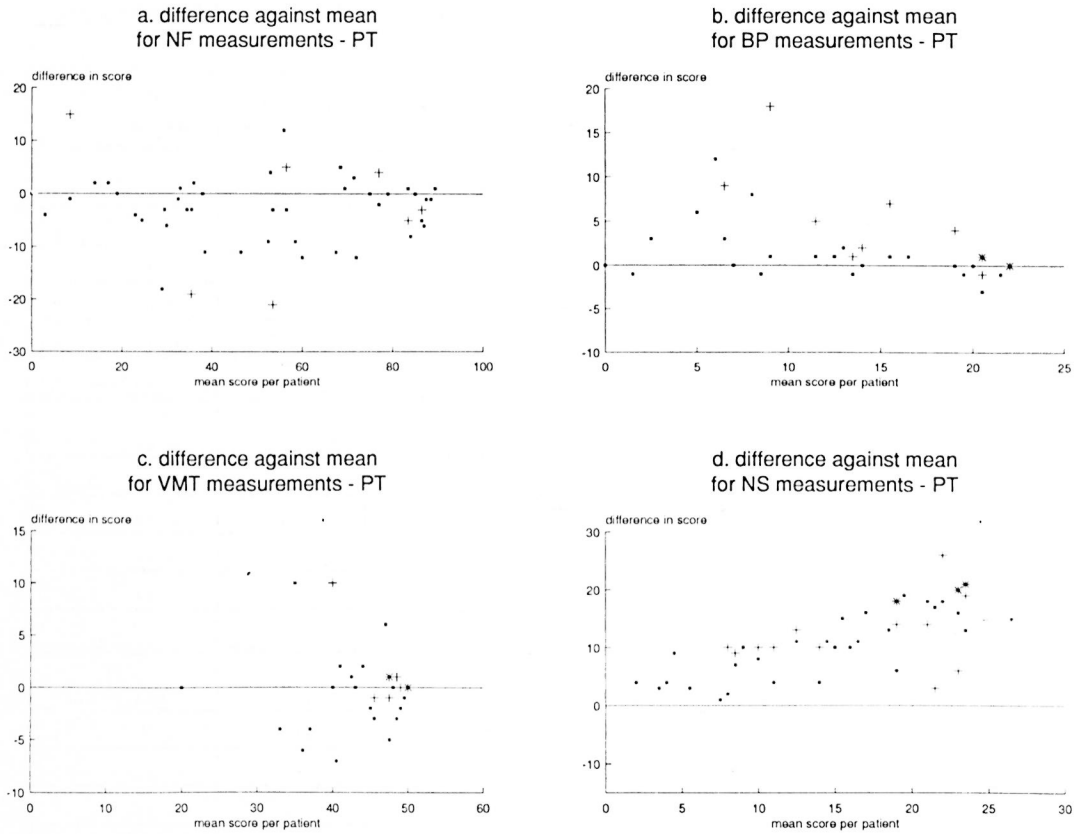


FIG. 1. Differences in scores against mean scores, per patient, for each test performed by the physio-technicians. Mean score for all patients is indicated by the solid line; + and * symbols represent several identical values.

there was evidence of a systematic difference between observers for all nerves tested ($p < 0.001$). Weighted κ values were very low, indicating that agreement between observers easily could be due to chance alone. Close examination of the results show that most variability was due to the "enlargement" component of this assessment.

Results of aggregate score per test. Figure 1 displays for each test the plot of the differences between the measurements obtained by the two examiners, against their mean. For the NF test (Fig. 1a) values were spread on each side of the zero value, with an excess toward the negative side, and the mean score difference was negative (Table 4). For the BP test, the differences in score spread more broadly toward the small mean score values, indicating that the variability of the measurements between examiners increased as the sensory function deteriorated. The level of precision of this test, therefore, appears to be related to the patient's mean score, indicating a poor reliability of this test in this population. For VMT, the mean measurement values appeared evenly distributed. The range of score differences was relatively narrow, and it can be calculated that this represents an expected variability of 12% over a total score of 50. Most mean difference values lie between 30 and 50 points, suggesting that motor function, as assessed by VMT, was quite homogenous in this population (Fig. 1c). Finally, for the clinical assessment of neuritis signs, the differences between measurements clearly increased with the mean score (i.e., with the severity of the clinical signs of neuritis), indicating that the variability of the responses obtained by the two observers was dependent upon the clinical status of the patient.

Health assistants

Description of data. Fifty other patients were examined by the two health assistants,

TABLE 3. *Results of nerve function assessment for all nerves by the health assistants.*

Nerve/test	No.	% Agreement	% Disagreement of 1 point	Mean difference	p[a]	Weighted κ	CI of weighted κ
NF							
r ulnar	50	42	12	-0.36	0.619	0.768	-[b]
l ulnar	50	44	12	-0.98	0.016	0.696	0.26 - 1.12
r median	50	36	10	-1.92	0.001	0.551	-0.10 - 1.20
l median	50	34	24	-1.20	0.006	0.616	0.02 - 1.21
r post-t	50	46	12	-0.32	0.581	0.744	0.37 - 1.12
l post-t	50	40	24	0	0.833	0.743	0.36 - 1.15
BP							
r ulnar	49	75	10	0.10	0.455	0.637	0.44 - 0.84
l ulnar	49	75	14	0.12	0.353	0.668	0.49 - 0.85
r median	49	82	16	0.12	0.148	0.616	0.39 - 0.84
l median	49	82	12	0.14	0.175	0.655	0.44 - 0.87
r post-t	50	68	14	0.20	0.284	0.704	0.34 - 1.07
l post-t	50	66	22	0.20	0.255	0.652	-
VMT							
r facial	49	78	16	-0.08	0.489	0.581	-0.93 - 2.08
l facial	49	65	28	-0.16	0.121	0.444	-0.85 - 1.73
r ulnar	49	75	20	-0.12	0.243	0.757	0.12 - 1.39
l ulnar	49	78	22	-0.06	0.548	0.809	0.26 - 1.35
r median	48	87	10	-0.10	0.187	0.627	-2.51 - 3.76
l median	48	90	8	-0.08	0.312	0.804	-0.71 - 2.31
r radial	48	92	8	-0.04	0.625	0.475	-7.96 - 8.92
l radial	48	94	4	-0.08	0.250	0.484	-7.88 - 8.84
r c per	48	98	2	0.02	1.000	0.948	-0.49 - 2.39
l c per	48	98	-	-0.06	1.000	0.657	-6.52 - 7.84
NS							
r ulnar	50	48	26	-0.54	0.001	0.121	-[b]
l ulnar	50	38	40	-0.46	0.001	0.129	-
r median	50	86	2	-0.56	0.015	0	-
l median	50	88	4	-0.42	0.031	0	-
r radial	49	73	6	-0.55	0.001	0	-
l radial	49	60	20	-0.60	0.001	0	-
r c per	49	62	34	0.06	0.860	0.227	-
l c per	49	68	26	0.06	0.469	0.267	-
r post-t	49	72	16	-0.30	0.025	0.153	-
l post-t	49	74	14	-0.36	0.007	0.163	-

[a] Tables 2 and 3 display mean differences together with the results (p value) of the Wilcoxon test, which is based on the median difference. Means rather than medians are displayed as the sign of the mean difference between observers was informative as to the direction of the bias.

[b] Standard error could not be calculated due to negative square root values.

TABLE 4. *Aggregate score per test: mean differences in score between physio-technicians for each test.*

Test	No.	Mean difference ± S.D.
NF	49	−2.92 ± 6.92
BP	49	1.61 ± 3.72
VMT	47	−0.28 ± 3.13
NS	46	11.07 ± 5.91

16 (30.6%) females and 34 (69.4%) males. Mean age was 36.2 years old; 29.2 years for females, 39.4 years for males. The distribution of leprosy by type was as follows: TT 1 (2%), BT 10 (20%), BT/BB 1 (2%), BB/BL 1 (2%), BL 23 (46%), LL 9 (18%), other or unknown 5 (10%). Twenty patients (48%) were in reaction and/or were experiencing neuritis. The age and sex distributions of this population were comparable to the population tested by the physio-technicians, but there was some difference in the distribution of leprosy types and in the number of patients with reaction and/or neuritis.

Results per test. *NF test.* The percent agreement in the NF test ranged from 34% to 46%. In 10% to 24% of the cases, the observers disagreed by 1 point over a total score of 15 points. Mean differences between measurements were negative for all nerves (except the left posterior tibial), showing that one observer was constantly rating higher than the other. There was significant evidence of systematic difference for the right and left median nerve and for the left ulnar nerve. Weighted κ statistics showed fair-to-good agreement better than expected by chance (0.616 to 0.768), but confidence intervals were wide.

BP test. There was good agreement in the BP test in general (66% to 82%), better in the hands than in the feet. Mean differences were positive for all nerves but showed no significant evidence of systematic difference ($p > 0.1$). Weighted κ statistics indicated fair-to-good agreement better than expected by chance.

VMT. Except for the facial and the ulnar nerves, there was a high percent agreement with VMT between observers for this test (87% to 98%) and there was no evidence of systematic difference between observers for all nerves ($p \geq 0.4$). Except for the facial and the radial nerves, $w\kappa$ statistics showed fair-to-good agreement better than expected by chance, but with wide confidence intervals, not excluding chance agreement.

NS. The agreement for clinical signs of neuritis was extremely variable, ranging from 38% to 88% and, as for the physio-technicians, $w\kappa$ values indicated that any agreement between examiners for these nerves was likely to be due to chance alone. Except for the common peroneal nerve, there was strong evidence of a systematic difference between the observers for all nerves ($p < 0.05$).

Results of aggregate score per test. When using NF, the health assistants showed a pattern of measurements similar to the physio-technicians. In Figure 2a, values were spread on each side of the zero value, with an excess toward the negative side and the mean score difference was negative. For the BP, the mean score values were evenly spread. For VMT, there was little variation in individual mean score values which appeared evenly distributed around the mean score value, ranging from 15 to 30. The limits of agreement were narrower than those obtained for the physio-technicians, which probably is due to the smaller scale used by the health assistants (0–30 instead of 0–50), offering less variability in the rating.

For NS (Fig. 2d), the variability of the measurements between observers increased with the severity of the clinical signs of neuritis, as already noted with the physio-technicians, but in the opposite direction.

DISCUSSION

The data collected by the physio-technicians and the health assistants come from different populations which have been tested at different periods and under slightly different conditions. They are, therefore, not strictly comparable, but patterns of variability concerning the two types of leprosy workers can be identified.

For the assessment of sensory function, although agreement between the physio-technicians when using the BP was apparently better than the NF method for all the nerves tested this is almost entirely due to the difference in scales, a smaller scale allowing less variability in the rating. Weighted kappa statistics make some allowance for scale difference and were quite similar for

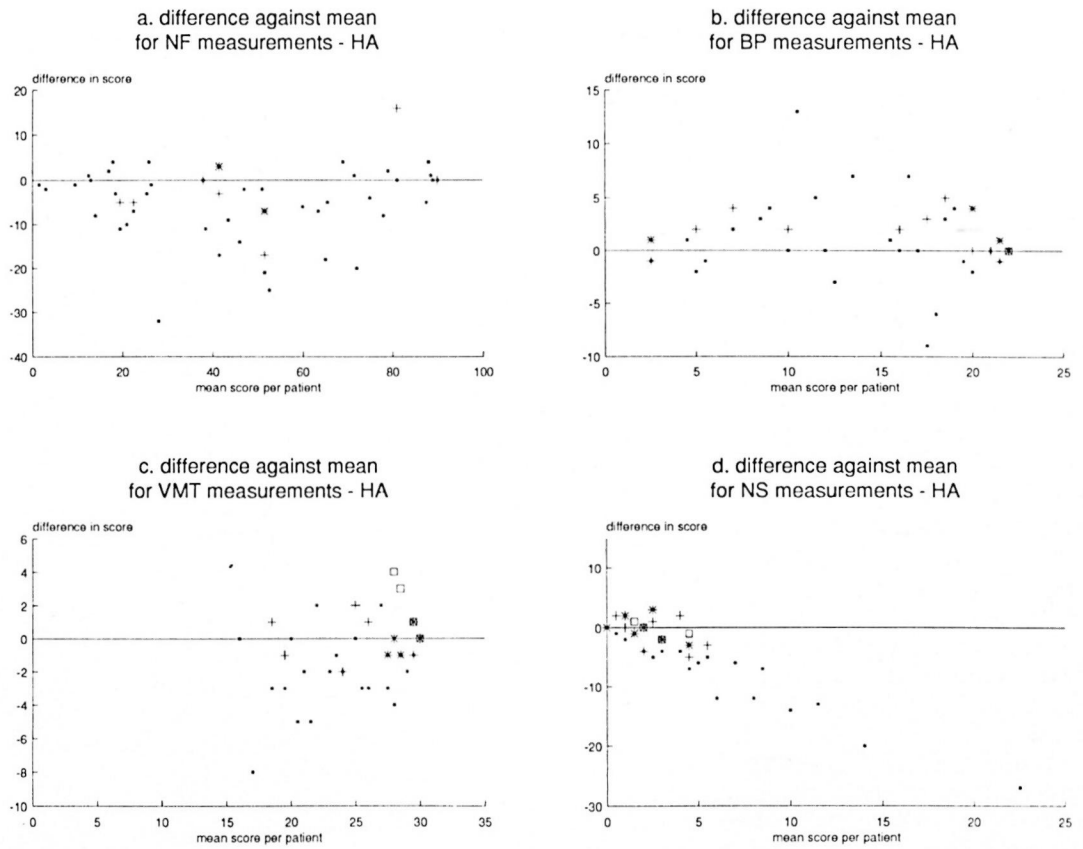


FIG. 2. Differences in scores against mean scores, per patient, for each test performed by the health assistants. Mean score for all patients is indicated by the solid line; +, * and □ symbols represent several identical values.

the two methods. In addition, the BP method seems poorly reliable in this population since the differences in aggregated scores appear to be dependent upon the mean score value, an effect not seen with the NF method. The presence of a systematic difference between observers in opposite directions when using the NF and BP methods reflect the subjectivity inherent in these tests, in the absence of clear standard reference values. The systematic difference between physio-technicians when using the BP method could be explained by their lack of experience with this test since they routinely assess sensory function with the graded NF rather than with a BP. On the other hand, the systematic difference observed between them when performing the NF test may reflect the technical difficulty of this test and, despite training, the subjectivity inherent in this measurement. Lastly, for the two methods the agreement between observers appeared better when examining the nerves of the hands than the nerves of the feet, which can be related to the relative difficulty of performing sensory tests in the feet compared to the hands ([18]).

For the health assistants, absolute agreement was better with the BP than with the NF method, although weighted kappa statistics were comparable. However, BP was not liable to a systematic difference between observers. This is not surprising, since health assistants routinely perform sensory testing in leprosy patients with a single nylon filament, thus testing an "all or nothing" response to a single stimulus, which is very similar to the BP method. They were taught to use the NF just before the study, and it is remarkable that their results are not so different from those obtained by the physio-technicians.

There is an apparent paradox in the assessment of motor function by the physio-

TABLE 5. *Aggregate score per test: mean differences in score between health assistants for each test.*

Test	No.	Mean difference ± S.D.
NF	50	−4.96 ± 8.56
BP	49	0.88 ± 3.35
VMT	48	−0.81 ± 2.13
NS	49	−3.31 ± 5.81

technicians and the health assistants: despite a high percent agreement, w_K statistics are extremely variable and do not exclude chance agreement (particularly for the facial and the radial nerves). In fact, for both groups of observers, cross-tabulations of single measurements show that similar ratings had been attributed to most of the patients by the two observers. Due to the extreme homogeneity of the ratings, agreement due to chance cannot be excluded, and this is further reflected in the wide confidence intervals. The high percent agreement between examiners observed in the study can thus be explained either by the homogeneity of the motor status of the patients or by the impossibility for the examiners to differentiate between the patients, due to their inability to apply the test properly or due to the inability of the test to differentiate between them. The aggregated score per test confirms the similarity among the patients since, except for one value, all mean score values lie between 32 and 50 (on a scale of 0–50) for the physio-technicians (Fig. 1c) and between 17 and 30 (on a scale of 0–30) for the health assistants (Fig. 2c). The high percent agreement thus could be due to chance alone, in a population with a high prevalence of "conserved" motor function (as assessed by VMT) and the small variability between observers probably reflects within-observer variation. The low agreement observed between physio-technicians and health assistants for the facial nerve probably reflects the technical difficulty in testing and grading motor function on the eyelids.

Concerning the assessment of clinical signs of neuritis, all results are consistent with the poor repeatability of the method, both by the physio-technicians and the health assistants. A strong systematic difference between observers is found for almost all nerves tested, and the variability of the measurements increases with the severity of the neuritis, reflecting the subjectivity of this method. Examining the three components of this assessment, it appears that most variability and systematic differences are related to the assessment of nerve enlargement. This shows the difficulty of relying on clinical examination alone to assess the presence and the severity of nerve damage and/or neuritis.

Sensory testing with the NF and BP methods and VMT are largely used throughout the world as means of assessing nerve function in patients with leprosy. Few studies have been carried out, however, on their repeatability with the same or different examiners. According to Bell-Krotoski and Buford [2], hand-held tests introduce major variation in instrument application force and frequency from examiner to examiner when there is no control of the force with which the instrument is applied. In this context, one of the main advantages of NF, as opposed to the BP, is that the bend of the filament provides some control on the application amplitude and on the vibrations exerted by the examiner [3]. Bell-Krotoski and Tomancik assessed the repeatability of graded NF and compared them to other hand-held instruments, including BP [3]. On repeated testing, they found that the range of forces of application with the BP varied broadly (in grams), whereas with the NF there was a small variation (in milligrams), and the BP compared more closely with the thickest NF (300 g). The authors concluded that if diameters and lengths are correct, the application forces of the filaments were repeatable within a predictable range. Conversely, the pressure stimulus applied on the skin with the BP appears difficult to standardize.

When assessing the repeatability of a measurement method, assuming that within-patient variability is randomly distributed, the following aspects must be considered: the instrument used to make the measurement (in this case, the instrument used to apply the stimulus), the observer who performs the test and records the measurement and the time (or occasion) of testing. In their study, Bell-Krotoski and Tomancik tested the repeatability of the stimulus created by the instrument but not the response to that stimulus. Their results suggested that

the stimulus created by the NF was more repeatable than that created by the BP ([3]). The forces of application of the repeated stimuli were measured very precisely but the stimuli were not applied on leprosy patients, and no information was given regarding the examiners (number, experience, quality), the time of testing and the time between repeated testings. As a complement to this investigation, we were interested, in our study, in evaluating how much variability could be expected with these (and other) tests, given their intrinsic characteristics, when applied successively by different examiners under field conditions.

Although NF design is said to control for the quality of the stimulus, we found an important variability between both types of observers with this test. This can be explained by the differences in the technique of application of the stimulus by the examiners (speed, length of time), by the subjective response of the patient to the stimulus (related to his/her personal threshold), and by the subjective assessment of the patient's response to the stimulus by the examiner, which is dependent upon training and experience ([2]). The different variability between examiners observed with both the NF and BP methods can be explained by the different scales used for grading the responses of the patients to the stimulus, an "all or nothing" response allowing much less variability of the measurements than a response graded on a larger scale. These elements of variability can explain why, despite a better control of the tactile stimulus with the NF, there is less consistency between health assistants when using NF rather than BP and why, despite training and experience, physio-technicians are not showing a much better agreement when using graded NF than BP. Since subjectivity of the patients cannot be controlled, all efforts should then concentrate on reducing the observer's subjectivity through on-going training and supervision in order to keep variability of the measurements to a minimum.

Concerning VMT, Naafs suggested that it was a reliable method for the assessment of nerve involvement, especially in patients presenting "severe nerve damage" ([19]) but the study was based on 12 patients tested at successive periods of time during treatment. In another study ([16]), it was found that VMT was not "reliably reproducible," but repeated testings were done 1 to 3 weeks apart on 20 patients under treatment.

The interpretation of w_κ often relies on whether the source of variation is the result of within- or between-observer disagreement. Extensive variation between observers can be partly explained by variation within an observer. In order to disentangle the respective effects of within- and between-observer variation, we could have asked the examiners to repeat their readings on each subject. Within-observer variation would have been estimated for each observer by calculating the percent agreement between the first and second readings and the mean value of the two readings would have been used to assess between-observer variation and to estimate the systematic difference ([7]). In this study, we chose not to do this since it was difficult to impose four nerve function tests per day on the patients.

In this study, the variability between observers in the assessment of nerve function was the least when applying tests that they use routinely, such as the VMT for the physio-technicians and the health assistants. Conversely, there was evidence of greater variability or systematic differences among examiners when using a test that they did not perform frequently (such as the BP for the physio-technicians), or when the test was liable to extensive judgmental variation, as was the case with the assessment of neuritis signs (NS), by both the physio-technicians and the health assistants. Agreement among observers was better with tests coding on small scales (such as the BP) than with tests coding on large scales (NF), which is not surprising but is detrimental to the fine assessment of nerve function and to the early detection of nerve function loss ([18]).

The variability and the presence of a systematic difference among observers when using the NF test shows the importance but also the limits of experience since, despite experience and proper training, there was evidence of a systematic difference between physio-technicians for some nerves. In addition to continuous training, this reinforces the need for close supervision of leprosy workers involved in nerve examination and regular quality control procedures in order to keep inter-observer variability to a min-

imum and to avoid misclassification of nerve function. It also would be suitable that the methods and scales of the various tests used for the assessment of nerve function in LCPs be standardized and unified. Finally, it could be interesting to compare the tests used to assess nerve function in other types of peripheral neuropathies (such as diabetes), and to investigate their potential application to leprosy.

SUMMARY

One of the major problems in leprosy is to detect any change in nerve function early enough so as to increase the chances of recovery and prevent disability. Several tests have been developed to assess nerve function and are used in leprosy control programs worldwide, but they are frequently performed by different workers on different occasions and under variable conditions. In this study we investigated the variability between different groups of observers in the assessment of nerve function in leprosy patients in Ethiopia. Sensory function was assessed by using a set of nylon monofilaments (NF) and a ball-point pen (BP), and motor function was assessed by using voluntary motor testing (VMT). We also studied the variability between observers in the assessment of the clinical signs of neuritis.

Duplicate measurements were performed in random order on 50 leprosy patients by two physio-technicians and on 50 other patients by two health assistants. The percent agreement between observers was calculated for each single nerve, and weighted kappa statistics were used to assess whether agreement was better than expected due to chance alone. Systematic differences between observers were evaluated using the Wilcoxon signed rank test. On sensory testing, inter-observer variability was found to be related to the training and experience of the observer, to the nerve tested, and to the neurological status of the patient.

When tests were performed by physio-technicians, we observed 32% to 58% agreement with the NF test and 71% to 84% agreement with the BP test, measured on different scales. After weighting for the scale difference, the agreement seemed comparable with these methods but the differences in measurements with the BP test were found to be dependent upon the neurological status of the patient. The variability between observers differed according to the nerve tested, and there was some evidence of systematic differences between observers with both methods.

When performed by the health assistants, agreement was between 34% and 46% with the NF and between 66% and 82% with the BP tests. After weighting for the scale difference, the agreement seemed comparable but the BP was not liable to the systematic differences seen in the NF results. These differences could be attributed to the differences in the experience of the workers with these tests.

With the VMT, small variability between observers was found for all nerves tested, except the facial nerve, when performed by both the physio-technicians and by the health assistants (72% to 98% agreement). Change agreement, however, could not be excluded since the ratings were extremely homogenous. The assessment of neuritis signs was extremely variable between observers (14% to 41% agreement), with evidence of a systematic difference between observers.

Implications of these findings are discussed with the view to improving comparability of the nerve function tests under field conditions for early detection of nerve damage in leprosy patients.

RESUMEN

Uno de los prinicpales problemas de la lepra, es la detección de cambios en la función nerviosa lo suficientement temprano como para intentar mejorar las posibilidades de recuperación y prevenir el desarrollo de incapacidades. Se han desarrollado varias pruebas para establecer la función nerviosa y son de uso rutinario en los programas de control de la lepra en todo el mundo. Estas pruebas, sin embargo, se llevan a cabo por investigadores con diferentes preparación y bajo condiciones muy variables. En este estudio realizado en Etiopia, se investigó la variabilidad entre diferentes grupos de observadores encargados de valorar la función nerviosa en pacientes con lepra. La función sensorial se midió usando un juego de monofilamentos de nylon (NF) y un bolígrafo (BP), y la función motora se estableció usando una prueba motora voluntaria (VMT). También se estudió la variabilidad entre los observadores al establecer los signos clínicos de neuritis.

Dos fisiotécnicos efectuaron mediciones, por duplicado, en 50 pacientes con lepra. En otros 50 pacientes, las mediciones fueron efectuadas por 2 asistentes de

salud. Se calculó el porcentaje de concordancia entre los observadores para cada uno de los nervios y se usó una prueba estadística para establecer si la concordancia fue mejor que lo esperado tan solo por azar. Las diferencias sistemáticas entre los observadores se evaluaron usando la prueba de rangos de Wilcoxon. Se encontró que en la prueba sensorial, la variabiliad entre los observadores estuvo relacionada con el entrenamiento y la experiencia del observador, con el nervio probado, y con el estado neurológico del paciente.

Cuando las pruebas se efectuaron por los fisiotecnicos se observó una concordancia del 32 al 58% con la prueba del NF y del 71 al 84% con la prueba BP, medidas en diferentes escalas. Después de las correcciones necesarias, la concordancia pareció comparable con estos dos métodos pero también se encontró que las diferencias en las mediciones con la prueba de BP fueron dependientes del estado neurológico del paciente. La variabilidad entre los observadores difirió según el nervio probado, y hubieron diferencias sistemáticas (consistentes) entre los observadores con ambos métodos.

Cuando se efectuaron por los asistentes de salud, la concordancia estuvo entre el 43 y el 46% con la prueba NF y entre el 66 y el 82% con la prueba BT. Después de corregir por la diferencia en escalas, la concordancia pareció comparable pero la prueba BP no estuvo relacionada con las diferencias sistemáticas observadas en los resultados con la prueba de NF. Estas diferencias pueden atribuirse a diferencias en la experiencia de los observadores con estas pruebas.

Con la prueba VMT se encontró pequeña variabilidad para todos los nervios probados, excepto para el nervio facial, cuando se efectuó tanto por los fisiotécnicos como por los asistentes de salud (concordancia del 72 al 98%). El establecimiento de los signos de neuritis fue extremadamente variable entre los observadores (concordancia del 14 al 41%) y las diferencias entre ellos fueron consistentes.

Se discuten las implicaciones de estos hallazgos con la idea de mejorar la comparabilidad de las pruebas de función nerviosa bajo condiciones de campo para la detección de daño nervioso temprano en los pacientes con lepra.

RÉSUMÉ

L'un des problèmes majeurs de la lèpre est la détection des modifications de la fonction nerveuse suffisamment tôt pour augmenter les chances de récupération et prévenir les incapacités. Différents tests ont été développés pour évaluer la fonction nerveuse et sont utilisés dans les programmes de lutte contre la lèpre à travers le monde, mais ils sont souvent exécutés par des personnes différentes à différentes occasions et dans des conditions variables. Dans cette étude, nous avons examiné la variabilité, entre différents groupes d'observateurs, de l'évaluation de la fonction nerveuse de malades de la lèpre en Ethiopie. La fonction sensorielle a été évaluée en utilisant des mono-filaments de nylon et la pointe d'un stylo à bille, et la fonction motrice par le test de motricité volontaire. Nous avons également étudié la variabilité de l'évaluation des signes cliniques de névrite parmi les observateurs. Des mesures doubles ont été réalisées dans un ordre aléatoire sur 50 malades de la lèpre par deux physio-techniciens et 50 autres malades par deux assistants sanitaires. Le pourcentage d'agrément entre les observateurs a été calculé pour chaque nerf, et un kappa pondéré a été utilisé au point de vue statistique pour évaluer si l'agrément était meilleur que celui attendu par le seul fruit du hasard. L'existence de différences systématiques entre les observateurs a été recherchée par le test de rang de Wilcoxon. Pour les tests sensoriels, on a trouvé que les différences inter-observateurs étaient associées à la formation et à l'expérience de l'observateur, au nerf testé et au statut nerologique du patient.

Quand les test étaient réalisés par des physio-techniciens, on a observé 32 à 58% d'agrément pour le test avec le filament de nylon et 71 à 84% d'agrément avec le test du stylo, mesurés sur des échelles différentes. Après pondération pour la différence d'échelle, l'agrément semblait comparable pour ces méthodes, mais on a trouvé que les différences de mesure avec le test du stylo dépendaient du statut neurologique du patient. La variabilité parmi les observateurs était différente selon le nerf testé, et il y avait des indices de différences systématiques parmi les observateurs avec les deux méthodes. Lorsque les tests étaient exécutés par les assistants sanitaires, l'agrément était entre 43% et 46% pour le test avec le filament de nylon, et entre 66% et 82% pour le test avec le stylo. Après pondération pour la différence d'échelle, l'agrément semblait comparable, mais le stylo n'était pas sujet aux différences systématiques observées dans les résultats des filaments de nylon. Ces différences pouvaient être attribuées à la différence d'expérience due personnel par rapport à ces tests. Avec le test moteur volontaire, on a trouvé une petite variabilité parmi les observateurs pour tous les nerfs testés, à l'exception du nerf facial, quand ils étaient exécutés aussi bien par les physio-techniciens que par les assistants sanitaires (72% à 98% d'agrément). On n'a pas pu, cependant, exclure un agrément par chance, du fait que les scores étaient extrêment homogènes. L'évaluation des signes de névrite était extrêmement variable parmi les observateurs (14% à 41% d'agrément), avec des signes évidents de différences systématiques entre observateurs.

Les implications de ces observations sont discutées avec comme objectif l'amélioration de la comparabilité des tests de la fonction nerveuse dans les conditions de terrain pour la détection précoce des altérations des nerfs chez les malades de la lèpre.

Acknowledgment. The authors would like to express their deep thanks to Mituko Kassa, Tafessa W/Mariam, Dejene Betemariam and Tesfaye Fanta, who kindly performed all these measurements in addition to their usual work, and to Dr. M. Waters for his useful comments. This study was made possible through the fi-

nancial contribution of LEPRA and other ILEP member/associations.

REFERENCES

1. BELL-KROTOSKI, J. A. Light touch deep pressure testing with Semmes Weinstein monofilaments. In: *Rehabilitation of the Hand.* 3rd edn. St. Louis: C. V. Mosby Co., 1990.
2. BELL-KROTOSKI, J. A. and BUFORD, W. L. The force/time relationship of clinically used sensory testing instruments. J. Hand Ther. **1** (1988) 76–85.
3. BELL-KROTOSKI, J. A. and TOMANCIK, E. The repeatability of testing with Semmes-Weinstein monofilaments. J. Hand Surg. **12** (1987) 155–161.
4. BECX-BLEUMINK, M., BERHE, D. and MAENNETJE, W. The management of nerve damage in leprosy control services. Lepr. Rev. **61** (1990) 1–11.
5. BLAND, J. M. and ALTMANN, D. G. Statistical methods for assessing agreement between two methods of clinical measurements. Lancet **1** (1986) 307–310.
6. BRANDSMA, W. Basic nerve function assessment in leprosy patients. Lepr. Rev. **52** (1981) 161–170.
7. BRENNAN, P. and SILMAN, A. Statistical methods for assessing observer variability in clinical measures. Br. Med. J. **304** (1992) 1491–1494.
8. BRYCESON, A. and PFALTZGRAFF, R. E. Leprosy. 3rd edn. Edinburgh: Churchill Livingstone, 1990.
9. DUNCAN, M. E. and PEARSON, J. M. H. Leprosy neuritis in pregnancy and lactation. Int. J. Lepr. **50** (1982) 31–38.
10. FLEISS, J. L. The measurement of interrater agreement. In: *Statistical Methods for Rates and Proportions.* New York: Wiley, 1981.
11. GOODWIN, C. S. The use of the voluntary muscle test in leprosy neuritis. Lepr. Rev. **39** (1968) 209–216.
12. HASTINGS, R. C., ed. *Leprosy.* Edinburgh: Churchill Livingstone, 1985.
13. JAMISON, D. G. Sensitivity testing as a means of differentiating the various forms of leprosy found in Nigeria. Lepr. Rev. **40** (1969) 17–20.
14. JOB, C. K. Nerve damage in leprosy. Int. J. Lepr. **57** (1989) 532–539.
15. KIRKWOOD, B. R. *Essentials of Medical Statistics.* Oxford: Blackwell Scientific Publications, 1988.
16. LEWIS, S. Reproducibility of sensory testing and voluntary motor testing in evaluating the treatment of acute neuritis in leprosy patients. Lepr. Rev. **54** (1983) 23–30.
17. LIENHARDT, C. and FINE, P. E. M. Type 1 reaction, neuritis and disability in leprosy: what is their current epidemiological situation? Lepr. Rev. **54** (1993) 9–33.
18. LIENHARDT, C., PASQUIER, R., LEMAITRE, C., BUTLIN, C. R. and WHEELER, J. Comparability of nylon filaments and ball pen in testing sensory function in patients with leprosy in Nepal and Ethiopia. (Abstract) Int. J. Lepr. **61** Suppl. (1993) 157A.
19. NAAFS, B. and DAGNE, T. Sensory testing: a sensitive method in the follow-up of nerve involvement. Int. J. Lepr. **45** (1977) 364–368.
20. NAAFS, B., PEARSON, J. M. H. and WHEATE, H. W. Reversal reaction: the prevention of permanent nerve damage; comparison of short and long-term steroid treatment. Int. J. Lepr. **47** (1979) 7–12.
21. NAAFS, B. and VAN DROEGENBROECK, J. B. A. Etude comparative d'une série de nerfs lépreux décomprimés chirurgicalement par rapport aux nerfs controlatéraux non opérés. Med. Trop. **37** (1977) 763–770.
22. PEARSON, J. M. H. The evaluation of nerve damage in leprosy. Lepr. Rev. **53** (1982) 119–130.
23. PEARSON, J. M. H. and ROSS, W. F. Nerve involvement in leprosy: pathology, differential diagnosis and principles of management. Lepr. Rev. **46** (1975) 199–212.
24. SEMMES, J., WEINSTEIN, S., GHENT, L. and TEUBER, H. *Somatosensory Changes After Penetrating Brain Wounds in Man.* Cambridge: Harvard University Press, 1960, pp. 4–41.
25. WATSON, J. M. *Preventing Disability in Leprosy Patients.* London: The Leprosy Mission International, 1986.
26. WATSON, J. M. Disability control in a leprosy control programme. Lepr. Rev. **60** (1989) 169–172.
27. WHO Study Group. Epidemiology of leprosy in relation to control. Geneva: World Health Organization, 1985. Tech. Rep. Ser. 716.

ANNEX 1

Assessment of clinical signs of neuritis (NS)

The three main signs and symptoms of neuritis checked by the examiners were: nerve pain, nerve tenderness, and nerve enlargement. The examiner asks the patient if he/she feels any spontaneous pain and, if yes, where it is located. Then the examiner checks for nerve tenderness (pain induced by palpation) and for nerve enlargement. The following nerves are examined: superficial radial, ulnar, median, lateral popliteal, posterior tibial (common peroneal).

The score is as follows: a. Nerve pain (spontaneous): no pain = 0, moderate pain = 1, severe pain (incapacitating) = 2. b. Nerve tenderness (induced): no tenderness = 0, moderate tenderness = 1, severe tenderness (withdraw) = 2. c. Nerve enlargement: not enlarged = 0, moderately enlarged = 1, markedly enlarged = 2.

The score for signs and symptoms of neu-

ritis is calculated for each nerve by adding up the results: pain + tenderness + enlargement. The highest score (6) is found when a patient has severe signs and symptoms of neuritis.

ANNEX 2

Sensory testing

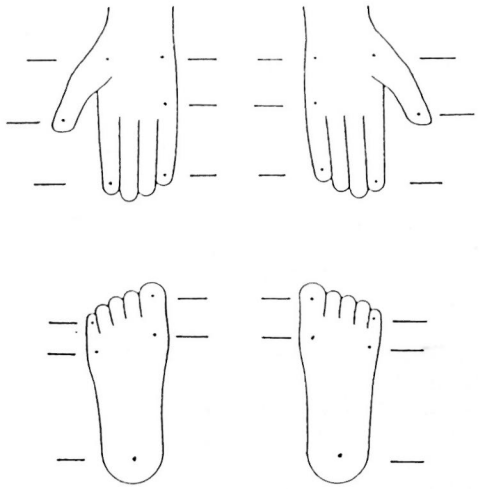
Nylon filaments. Five Semmes-Weinstein graded nylon monofilaments were used on specific sites of hands and feet (see figure). Each filament is applied slowly to bending, held for 1 to 2 seconds, and lifted slowly while the patient's eyes were closed. Each filament is applied three times in each tested area. Each time, the patient was asked to point to the stimulated area. If the patient pointed at least two times within 2 cm of the stimulated point, the response was correct for that filament and for that area of stimulation. The lightest filament (number 5) was applied first. If it was felt, the number 5 was recorded in the blank corresponding to the touched area. If not felt, the next heavier filament (number 4) was tried, and so forth for the remaining filaments. If no filament was felt, a zero was placed in the blank, showing complete anesthesia in this area.

The tested areas in the hand are supplied by the ulnar and the median nerves, and the area in the foot is supplied by the posterior tibial nerve. Scoring for sensory function applies only to those three nerves. The score per nerve is obtained by adding the results in the blanks corresponding to this nerve (see figure). The total score per nerve is 0 to 15.

The following nylon filaments are used: Hands: No. 5 = 0.2 g, No. 4 = 2.0 g, No. 3 = 4.0 g, No. 2 = 10.0 g, No. 1 = 300.0 g. Feet: No. 3 = 2.0 g, No. 2 = 10.0 g, No. 1 = 300.0 g.

Ball-point pen (adapted from Watson[25]). A ball-point pen is applied on specific sites of hands and feet, allowing a denting of 1 mm during 2 seconds, while the patient's eyes were closed. The patient was asked to point to the stimulated area with a finger. The ball-point pen was applied three times on each site. If the patient responded to at least 2 out of the 3 applications within 2 cm on a specific site, the response was correct and coded 1 for this site. An incorrect response was coded 0. The score per nerve was obtained by adding responses for each nerve (0 to 3 for ulnar and median nerve, 0 to 5 for posterior tibial nerve).

Sensory Testing in Hands and Feet



Nerve	Right	Left	Total
Ulnar			
Median			
Posterior tibial			

ANNEX 3

Voluntary motor testing

We used in this study an abridged version of the VMT, as proposed by Brandsma ([6]). The examiner first demonstrated the correct movement to the patient, and then asked him/her to repeat the movement spontaneously. If the patient was able to perform the full range of the demonstrated movement, he was asked to hold it against resistance and the results were coded accordingly.

When the examiner was a physio-technician, the movement was graded on the MRC scale (Medical Research Council. Aids to the investigation of peripheral nerve in-

juries. MRC Memo No. 7; London, 1943, HMSO.):

Grade 5 = Full range of movement against resistance; Grade 4 = full range of movement but less than normal resistance; Grade 3 = full range of movement but no resistance; Grade 2 = partial range of movement with no resistance; Grade 1 = perceptible contraction of the muscle not resulting in joint movement; Grade 0 = complete paralysis.

When the examiner was a health assistant, the movement was graded on the "SRMP" scale ([25]):

Grade 3 = Full range of movement against resistance (Strong); Grade 2 = reduced range of movement against resistance (Resistance reduced); Grade 1 = range of spontaneous movement reduced (Movement reduced); Grade 0 = no spontaneous movement (Paralysis).

The following nerves were tested:

Facial nerve: The patient was asked to close his/her eyes lightly, then tightly, and the strength of closure was scored on the above scale.

Ulnar nerve: The patient was asked to move his/her little finger straight out and a little up (abduction). Resistance was applied at the base of the finger if his movement could not be completed.

Median nerve: Abduction of the thumb; the patient was asked to move his thumb away from the palm of the hand at right angle. Resistance to this movement was applied perpendicular to the palm of the hand at the base of the thumb.

Radial nerve: Wrist extension; the patient was asked to move his/her clenched fist up. The examiner applied resistance trying to push the wrist down.

Lateral popliteal (common peroneal) nerve: Dorsiflexion of foot; the patient was asked to lift his/her foot. Resistance was applied by trying to push the foot down.

ANNEX 4

Unweighted kappa statistic

The unweighted kappa statistic is calculated as:

$$(P_o - P_e)/(1 - P_e),$$

where P_o is the proportion of measurements on which there is agreement between the two observers and P_e the proportion of measurements on which agreement would be expected to occur by chance alone ([10]).

Since, in this study, we were collecting measurements with multiple categories, opportunities for disagreement increased with the number of categories, and the variability between observers was likely to increase as well. In order to adjust for the seriousness of different levels of agreement, we calculated weighted kappa statistics, assigning increasing weights to increasing levels of disagreement. These weights were calculated objectively, using the "quadratic weights" proposed by Fleiss and Cohen ([10]).