Lepr Rev (1998) **69,** 257–266

Intertester reliability of manual muscle strength testing in leprosy patients

J. WIM BRANDSMA,*† WIM H. VAN BRAKEL,‡ ALISON M. ANDERSON,‡ ALEX J. KORTENDIJK,# KHADGA S. GURUNG & SHAHA K. SUNWAR‡

**Academic Medical Center, Amsterdam, †National Institute of Allied Health Professions, PO Box 1161, 3800 BD, Amersfoort, The Netherlands, ‡Green Pastures Hospital, PO Box 5, Pokhara, Nepal and #Aalberselaan 140, Amersfoort, The Netherlands*

Accepted for publication 15 June 1998

Summary This study reports the results of a study on the intertester reliability of manual muscle strength testing in leprosy patients with confirmed motor function loss of at least one nerve. Three testers graded the muscle strength of 72 patients in random order. Both hands and feet were graded. Strength was graded on a modified Medical Research Council Scale (9 points, 5, 4+, 4, 3+, 3, 2+, 2, 1, 0). The following movements were tested for strength: little finger and index finger abduction, intrinsic position of all four fingers, thumb abduction and opposition, foot dorsiflexion and eversion and extension of the big toe. The weighted kappa statistic was used to calculate the chance-corrected percentage of agreement between observers. Overall agreement for each of the 11 tests appeared to be good or very good (0·61–1·00). However, when data for hands or feet with normal strength or complete paralysis were excluded from the analysis, the reliability of the remaining mid-range scale was not acceptable (kappa 0·55–0·88, direct agreement range 11–41%). While the reliability of this scale could possibly be improved by special training, we feel that, for the evaluation of nerve function for leprosy patients with (suspected) nerve function loss, the extended 9-point VMT scale should only be used when direct intra- or intertester agreement is more than 80%.

Introduction

Manual muscle strength testing (MMST), commonly known as voluntary muscle testing (VMT) in leprosy, is an important technique in the assessment and evaluation of (motor) nerve (dys)function. Treatment decisions are often based on, and guided by the results of nerve function assessments, taking into consideration factors such as duration and severity of nerve function loss, clinical activity of the disease and findings on nerve palpation. A

Correspondence to: J. W. Brandsma, c/o The Leprosy Mission International, 80 Windmill Road, Brentford, Middlesex TW8 0QH, UK.

This study has been supported by a grant from the Q-M Gastmann-Wichers Foundation in the Netherlands.

0305-7518/98/069257+10 $1.00

Table 1. Six-point scale for interobserver reliability

Medical Research Council (MRC) scale[a] 6 grades	Modifications 9 grades
5 Full range of motion; full resistance	4+ moderate resistance
4 Full range of motion; some resistance	3+ minimal resistance
3 Full range of motion; no resistance	2+ nearly full range
2 Decreased range of motion	
1 Muscle flicker	
0 Complete paralysis	

[a] The MRC scale originally grades 4–5 against gravity and 0–3 with gravity eliminated. For muscle grading of small muscles, the effect of gravity is negligible.

timely diagnosis of the disease and decreasing nerve function followed by appropriate action may prevent the development of impairments and disabilities.[1,2]

Muscle strength testing is used in leprosy for diagnostic purposes, in studies to assess and evaluate the efficacy of medical and surgical interventions, and in epidemiological studies. The motor nerves commonly affected in leprosy are the facial, ulnar, median, radial, common peroneal and posterior tibial nerves.

Lienhart *et al.*[3] reported on the interobserver reliability of MMST of one muscle test for each of the nerves that can become paralysed in leprosy. Two physiotherapy technicians graded muscle strength on a 6-point scale and two field workers on a 4-point scale. Brandsma *et al.*[4] assessed the intra- and interobserver reliability for some intrinsic muscles and movements of the hand in nine tests, in 27 leprosy patients with confirmed loss of ulnar or median nerves. Reliability was assessed on a 6-point scale (Table 1). Results of the above studies showed acceptable reliability coefficients. Two reasons led us to study further the reliability of manual muscle testing. First, it would be desirable to know if acceptable reliability can be maintained if the muscle strength grading scale is refined. This could have important implications for diagnostic and management purposes. Second, the reliability coefficients of the muscle strength tests are influenced by the high proportion of patients with normal or with completely paralysed muscles which are relatively easy to grade. How would reliability be affected if the normal and completely paralysed muscles were left out of the analysis? The purpose of this study, therefore, was to establish the intertester reliability of muscle strength of muscles innervated by the ulnar, median and lateral popliteal nerves on a 9-point scale in a large population of leprosy patients.

Methods

The study was conducted at the Green Pastures Hospital in Pokhara, Nepal. This 100-bed hospital is a referral hospital for leprosy patients in the Western Region of Nepal. For this study, both in- and out-patients were used. Informed consent was obtained from all patients. Seventy-two patients were admitted to the study.

Both hands and feet were tested. A hand or foot was not tested if there was a severe

Table 2. Nine-point scale used to grade muscle strength

Nerve	Movement tested
Ulnar	abduction little finger intrinsic position fingers (4) abduction index finger
Median	abduction thumb opposition thumb
Lateral popliteal	dorsiflexion foot eversion foot extension big toe

deformity or if testing was painful, which could compromise the test results. The tests have been described elsewhere by the first author.[5]

Intertester reliability was assessed between three testers, two experienced technicians from the hospital (5 and 10 years experience in manual muscle strength testing), and a visiting physiotherapist with more than 20 years experience in nerve function testing in leprosy patients. A 9-point scale was used to grade strength of muscles innervated by the ulnar, median and lateral popliteal nerves (Tables 1 and 2).

Prior to the start of the study, the assessors had one session to discuss the testing protocol. There was no practice session with patients prior to the study, to discover and discuss possible differences in interpretation of grades. Patients were tested in random order by the assessors, who had no information on test results of the other assessors. Testing for each patient was usually completed within 1 h.

STATISTICAL METHODS

Data were entered in Epi Info version 6, an integrated software program developed by the Centres for Disease Control in the United States and the World Health Organization.[6] Paired observer agreement was evaluated using the percentage of direct agreement (the percentage of data pairs that agreed exactly) and the weighted kappa statistic. Weighted kappa is a coefficient of agreement between observers for categorical scales of more than two categories.[7] It may be interpreted as the percentage of agreement between the observers corrected for chance and taking into account scaled disagreement. Quadratic disagreement weights were used.[8] Ninety-five percent confidence intervals (95% CI) are presented for all weighted kappa values.

Data pairs of the right and left hand (or foot) were considered statistically independent for the purpose of this study. The 72 study subjects therefore contributed up to 144 sets of three data pairs. However, data from patients for whom one of the three testers did not perform a test were discarded. Because very few patients were scored as '1', these scores were recoded as '0'. This approach is clinically acceptable, as there is very little functional difference between these grades. For analysis purposes, therefore, the remaining scale was 8-point instead of 9-point. To examine the performance of the scale in the middle range (4+ to 2), a separate analysis was done after omitting the results of hands or feet scored by all three testers as '5' (normal strength) or '0' (paralysed).

Results

Table 3 shows the results of all the individual muscle strength tests using the data from the 9-point scale. While the weighted kappa values were mostly ≥0·80 (very good) for all three tester pairs, the direct agreement was much less good (range 40–84%). Agreement in the 'mid-range scale' (data from hands or feet scored 'normal' or 'completely paralysed' by all testers omitted) for six commonly tested movements is shown in Table 4. A dramatic drop is seen in both weighted kappa values and percent direct agreement (range 11–41%) for almost all sites and all tester pairs. Four typical agreement matrices are shown in Tables 5–8. Except in the upper left hand corner cells (0/0 or completely paralysed) and in the bottom right hand corner cells (5/5 or normal), very few data points lie on the diagonals. Omitting the data on both ends of the scale from the analysis, a marked reduction of the agreement indices is observed for the remaining scale. The difference between the full-scale agreement and the mid-scale agreement is illustrated in Figures 1 and 2.

Table 3. Intertester agreement of manual muscle strength testing (VMT) on a 9-point scale

Movement	Testers	Number of sides tested	Direct agreement (%)	Weighted kappa	95% CI
Abduction little finger	A, B	126	48	0·89	0·71–1
	A, C		44	0·86	0·68–1
	B, C		40	0·84	0·66–1
Intrinsic position index finger	A, B	120	51	0·9	0·72–1
	A, C		48	0·91	0·73–1
	B, C		52	0·88	0·70–1
Intrinsic position middle finger	A, B	123	66	0·94	0·76–1
	A, C		71	0·94	0·76–1
	B, C		66	0·94	0·75–1
Intrinsic position ring finger	A, B	121	63	0·93	0·75–1
	A, C		68	0·94	0·76–1
	B, C		63	0·91	0·73–1
Intrinsic position little finger	A, B	123	48	0·92	0·74–1
	A, C		45	0·9	0·72–1
	B, C		45	0·89	0·71–1
Abduction index finger	A, B	127	63	0·96	0·79–1
	A, C		58	0·94	0·77–1
	B, C		53	0·92	0·78–1
Abduction thumb	A, B	133	47	0·78	0·62–0·94
	A, C		56	0·82	0·66–0·98
	B, C		52	0·65	0·49–0·81
Opposition thumb	A, B	133	61	0·87	0·69–1
	A, C		59	0·9	0·72–1
	B, C		60	0·81	0·65–0·97
Dorsiflexion foot	A, B	124	78	0·94	0·77–1
	A, C		78	0·9	0·73–1
	B, C		73	0·84	0·67–1
Eversion foot	A, B	123	87	0·96	0·78–1
	A, C		84	0·87	0·69–1
	B, C		82	0·85	0·67–1
Extension big toe	A, B	112	55	0·91	0·73–1
	A, C		55	0·86	0·68–1
	B, C		46	0·81	0·63–0·99

Table 4. Intertester agreement of manual muscle strength testing on a 9-point scale: results after omitting hands and feet scored by all three testers as 'normal' or 'paralysed'

Movement	Testers	Number of sides tested	Direct agreement (%)	Weighted kappa	95% CI
Abduction little finger	A, B	92	30	0·76	0·56–0·96
	A, C		24	0·65	0·45–0·85
	B, C		17	0·62	0·42–0·82
Intrinsic position ring finger	A, B	66	32	0·67	0·43–0·91
	A, C		41	0·78	0·54–1
	B, C		32	0·64	0·42–0·86
Abduction index finger	A, B	72	35	0·88	0·64–1
	A, C		26	0·84	0·62–1
	B, C		17	0·77	0·55–0·99
Abduction thumb	A, B	79	11	0·49	0·30–0·68
	A, C		25	0·6	0·41–0·79
	B, C		19	0·4	0·6–0·54
Dorsiflexion foot	A, B	38	29	0·82	0·51–1
	A, C		29	0·66	0·37–0·95
	B, C		13	0·51	0·24–0·78
Eversion foot	A, B	28	39	0·88	0·49–1
	A, C		23	0·58	0·25–0·91
	B, C		15	0·55	0·22–0·88

Tables 5–7 show an asymmetrical distribution of the off-diagonal data points, indicating a systematic bias in the grading between the two testers, i.e. one tester is usually grading higher than the other tester. This bias occurred for each of the three tester pairs, only for some of the tests (data not shown). No consistent pattern could be found.

Discussion

The results of this study highlight a methodological problem that is not uncommon in studies of repeatability of measurements using a categorical scale. Because it is often not too difficult

Table 5. Agreement matrix of tester B and tester C for abduction of the little finger on a 9-point VMT scale. Direct agreement 39·7%; weighted kappa 0·84. Omitting normal (5/5) and paralysed (0/0) subjects, the direct agreement is only 17·4% and weighted kappa 0·62

Category	0	1	2	2+	3	3+	4	4+	5	Total
0	30		3	2	1	1				37
1			1							1
2	1		1		1	1	1			5
2+				2	3	1			1	7
3				1		3	3		2	9
3+					4	6	21	5	3	39
4						1	2		5	8
4+						1			1	2
5							5	5	10	20
Total	31	0	5	5	9	14	32	10	22	128

Table 6. Agreement matrix of tester A and tester B for abduction of the thumb on a 9-point VMT scale. Direct agreement 47·4%; weighted kappa 0·78. Omitting normal (5/5) and paralysed (0/0) subjects, the direct agreement is only 11·4% and weighted kappa 0·49

Category	0	1	2	2+	3	3+	4	4+	5	Total
0	9									9
1										0
2	1	1								2
2+	1		1	1						3
3			2	2	3	2				9
3+			2	4	3	3		2		14
4						1	1	3	4	9
4+			1		3	6	2		9	21
5				1	1	10	2	6	46	66
Total	11	1	6	8	10	22	5	11	59	133

Table 7. Agreement matrix of tester B and tester C for dorsiflexion of the foot on a 9-point VMT scale. Direct agreement 73·4%; weighted kappa 0·84. Omitting normal (5/5) and paralysed (0/0) subjects, the direct agreement is only 13·2% and weighted kappa 0·51

Category	0	1	2	2+	3	3+	4	4+	5	Total
0	6									6
1			1			1				2
2						1				1
2+			1			1	3			5
3							1			1
3+							1	1	4	6
4										0
4+							1	1	9	11
5						3		4	84	91
Total	6	0	2	0	0	7	6	6	97	124

Table 8. Agreement matrix of tester A and tester B for abduction of the index finger on a 9-point VMT scale. Direct agreement 48·4%; weighted kappa 0·89. Omitting normal (5/5) and paralysed (0/0) subjects, the direct agreement is 30·4% and weighted kappa 0·76

Category	0	1	2	2+	3	3+	4	4+	5	Total
0	22									23
1		1								1
2	2		2	1						5
2+		1	1	2	1	1				6
3			1	2	2	3				8
3+						3	1			4
4						6	0	2	2	10
4+							7	4	9	20
5						2	2	3	44	51
Total	24	2	4	5	3	15	10	9	55	127

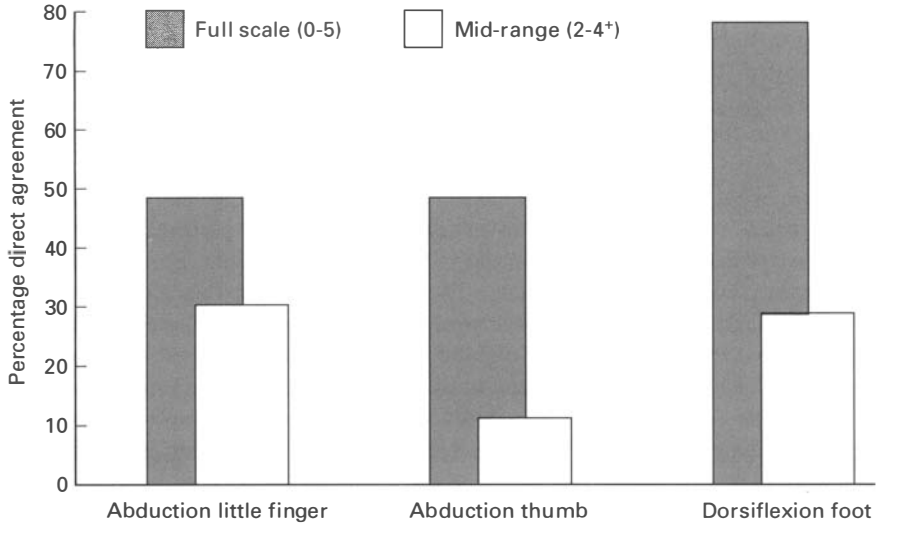

Figure 1. Differences between direct intertester agreement on the full 9-point VMT scale and on the 'mid-range' scale (hands and feet scoring normal or completely paralysed omitted).

to tell whether a test is completely normal or completely abnormal, agreement between testers tends to be good on both extremes of the scale. If there are many normal and completely abnormal test results, these will have a disproportionate effect on the agreement indicators that are calculated.

Between-tester agreement using the full 9-point scale appears good (Table 3), although the relatively low direct agreement is striking compared to the high values of the weighted kappas. No absolute standard exists for assessing kappa values, but many investigators use

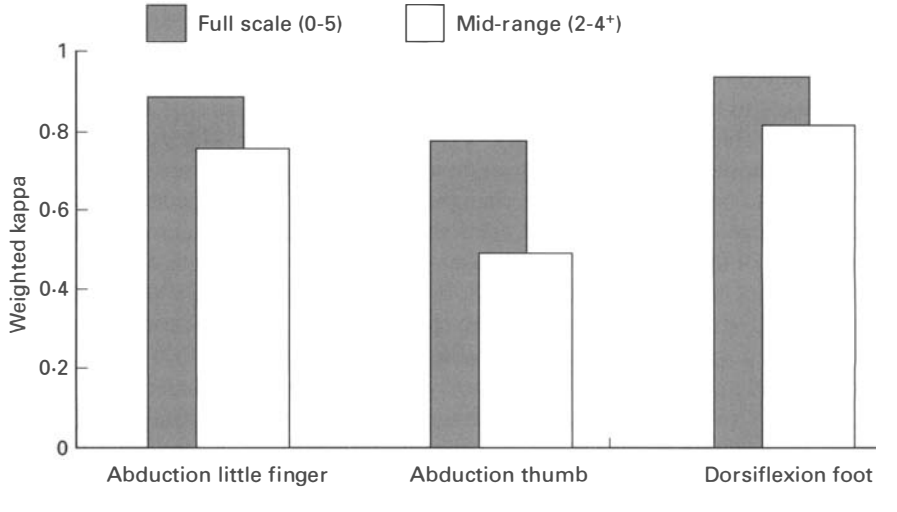

Figure 2. Differences between intertester agreement (weighted kappa) on the full 9-point VMT scale and on the 'mid-range' scale (hands and feet scoring normal or completely paralysed omitted).

the classification suggested by Altman: kappa 0·61–0·80: good; kappa 0·81–1·00: very good.[8] Using this classification, agreement between testers appears to be very good for most movements tested. Taken at face value, one would conclude that the 9-point VMT scale has very good reliability and can therefore be recommended for clinical use.

However, the clinical use of a rating scale such as the VMT is particularly important for patients who have some nerve dysfunction. Therefore, the performance of the scale in the mid-range, i.e. when there is muscle weakness, is at least as important as its ability to rate nerves as 'normal' or 'paralysed'. To investigate the mid-range performance of the 9-point VMT scale, we reanalysed the results omitting the data from muscle (groups) that were rated normal or paralysed by all three testers. The conclusion from this analysis (Table 4) is different. Direct agreement was very low, while the kappa values also dropped considerably. It should be noted that the value of weighted kappa is dependent on the prevalence of the individual grades and the number of categories of the scale. A decrease in the value of kappa after trimming the ends of the scale is therefore not unexpected. However, the size of the decrease and its concurrence with a marked decrease in direct agreement show that this is not due to an artifact.

The sample of agreement matrices shown in Tables 5–8 reveals a wide scatter of data points, with often more than one category difference between the testers. An exception is the rating of abduction of the index finger (Table 8), where the scatter is much less. Nevertheless, the mid-range direct agreement for this test was only 17·4–30·4%, depending on the tester pair. There was evidence of systematic bias between the testers at least for some of the tests. This could not be traced back to one particular tester or one or more specific tests. It is likely that specific training addressing these differences would result in improved reliability.

The benefit of the 9-point scale over the conventionally used 6-point scale is that it has more categories in the mid-range and is therefore potentially more sensitive than shorter scales. However, the current study shows that the rating of muscle strength in this mid-range is particularly unreliable. We conclude that the use of such an extended VMT scale should not be recommended, unless perhaps after special training reliability (direct inter- or intratester agreement) is more than 80%. It should be noted that the current results refer to inter-tester repeatability. It is likely that intra-tester repeatability would be better, but this should be investigated in a separate study.

A disadvantage of manual testing is that muscle strength is evaluated against what the examiner deems to be normal for a particular patient. This means that much experience is required for reliable testing, especially when children or the elderly are involved. To overcome this problem, the use of dynamometers has been recommended. Dynamometry has been shown to be more sensitive to changes in the MRC 3–5 range than MMST.[9–12] The grip and pinch dynamometers for the hand show excellent instrument reliability.[13,14] The additional value of dynamometers in the assessment of muscle strength of the hand and foot in leprosy patients needs to be investigated. In the 3–5 range on the VMT scale, the values obtained with a dynamometer could confirm changes that are also noticed with the VMT, or could indicate changes that are not noticeable on manual testing. VMT and dynamometry could be assessed and recorded simultaneously, as suggested by Brandsma.[15]

It is important to realize that in strength testing in the so-called 'lumbrical position', the interosseus muscles are tested rather than the lumbricals. This is especially important for the index and middle fingers, which, in more than 95% of patients with an ulnar palsy, will be weak or may even 'claw'.[16]

At the outset of the study, it was not planned to test big toe extension (extensor hallucis

longus, EHL). Big toe extension, however, is routinely assessed at the Green Pastures Hospital and the two physiotechnicians who participated in the study indicated that they often recorded isolated weakness of big toe extension, with normal strength of eversion and dorsiflexion. It was therefore decided to include this test in the study. Isolated weakness of EHL was observed in 26 feet (22 patients), which is 23% of all feet tested. Is it possible to have isolated damage of the nerve branch supplying the EHL? Could it be that weakness in the EHL precedes weakness in dorsiflexion and eversion? An answer to these questions needs to be obtained through a prospective study.

Fritschi, Jehin and Palande recommend the use of so-called functional muscle tests in the field, e.g. thumb to little finger opposition.[17–19] It should be realized that these tests may still be negative when 50–70% of muscle strength has already been lost.[20] They are only of (limited) value when patients or suspected patients have to be screened in large numbers or when competency and quality of the leprosy staff does not allow for grading of muscle strength.

In conclusion, in this study, the mid-scale range (4+ to 2) of the 9-point VMT scale showed unacceptable intertester reliability. Therefore, we recommend that the 9-point scale (or any further sophistication, such as an 11-point scale) in the assessment of nerve function in leprosy patients is only used when acceptable direct intra- or intertester agreement has been obtained.

Caution is needed when basing conclusions of reliability studies on weighted kappa values alone. This is particularly true if the study population contains a high proportion of normal or completely abnormal test results. Direct agreement, reliability of the mid-scale range and examination of agreement matrices can give essential extra information.

The use of dynamometry as an adjunct to manual muscle testing in leprosy should be investigated.

Acknowledgements

The help and support of Dr Frauke C. Wörpel, Superintendent of Green Pastures Hospital, and of other staff at the hospital is gratefully acknowledged. The work at the hospital is dedicated to the glory of God.

References

[1] Brandsma JW, Heerkens YF, Lakerveld-Heyl K. The International Classification of Impairments, Disabilities, and Handicaps in leprosy control projects. *Lepr Rev*, 1992; **63:** 337–344.

[2] Brandsma JW. Terminology in leprosy rehabilitation and guidelines for nerve function assessment. *Trop Geograph Med*, 1994; **46:** 98–102.

[3] Lienhart C, Currie H, Wheeler JG. Inter-observer variability in the assessment of nerve function in leprosy in Ethiopia. *Int J Lepr*, 1995; **63:** 62–76.

[4] Brandsma JW, Schreuders T, Birke JA *et al.* Manual muscle strength testing. Intra- and inter-observer reliability of the intrinsic muscles of the hand. *J Hand Ther*, 1995; **8:** 185–190.

[5] Brandsma JW. Basic nerve function assessment in leprosy patients. *Lepr Rev*, 1981; **52:** 161–170.

[6] Epi Info, version 6. *A word processing, database, and statistics system for epidemiology on microcomputers.* US Department of Health & Human Services, 1994.

[7] Cohen J. Weighted kappa: nominal scale agreement with provision for scalded disagreement or partial credit. *Psych Bull* 1968; **70:** 213–220.

[8] Altman DG. *Practical statistics.* Chapman & Hall, London, New York, 1991: pp. 404.

[9] Bohannon RW. Manual muscle test scores and dynamometer test scores of knee extension strength. *Arch Phys Med Rehab*, 1986; **67:** 390–392.

[10] Schwartz S, Cohen ME, Herbison G, Shah A. Relationship between two measures of upper extremity strength: manual muscle test compared to hand-held myometry. *Arch Phys Med Rehab*, 1992; **73:** 1063–1068.
[11] Aitkens S, Lord J, Bernauer E *et al*. Relationship of manual muscle testing to objective strength measurements. *Muscle Nerve*, 1989; **12:** 173–177.
[12] Grifin JW, McClure MH, Bertorini TE. Sequential isokinetic and manual muscle testing in patients with neuromuscular disease. A pilot study. *Phys Ther*, 1986; **66:** 32–35.
[13] Hamilton GF, McDonald C, Chenier TC. Measurement of grip strength. Reliability of the sphygmomanometer and Jamar grip dynamometer. *J Orthop Sports Phys Ther*, 1992; **16:** 215–219.
[14] Mathiowetz V. Reliability and validity of grip and pinch strength measurements. *Phys Rehab Med*, 1991; **2:** 201–212.
[15] Brandsma JW. Manual muscle strength testing and dynamometry for bilateral ulnar neuropraxia in a surgeon. *J Hand Ther*, 1995; **8:** 191–194.
[16] Brandsma JW. *Intrinsic/Minus/Hand. (Patho)kinesiology, rehabilitation and reconstruction.* PhD Dissertation, 1993. University of Utrecht, The Netherlands.
[17] Fritschi EP. Field detection of early neuritis in leprosy. *Lepr Rev*, 1987; **58:** 173–177.
[18] Jehin EG, Smith WCS, Day R. A quick VMT and ST for the hands compared to a standard VMT and ST. *14th Int Lepr Congress*, Orlando, 1993.
[19] Palande D, Bowden REM. Early detection of damage to nerves in leprosy. *Lepr Rev*, 1992; **63:** 60–72.
[20] Ploeg van der RJO, Oosterhuis HJGH, Reuvekamp J. Measuring muscle strength. *J Neurol*, 1984; **231:** 200–203.