Published in final edited form as:
Intensive Care Med. 2010 June ; 36(6): 1038–1043. doi:10.1007/s00134-010-1796-6.

Inter-rater reliability of manual muscle strength testing in ICU survivors and simulated patients

Eddy Fan,
Division of Pulmonary and Critical Care Medicine, Johns Hopkins University, 1830 East Monument Street, 5th Floor, Baltimore, MD 21205, USA, Tel.: +1-410-9553467, Fax: +1-410-9550036

Nancy D. Ciesla,
Department of Physical Medicine and Rehabilitation, Johns Hopkins Hospital, Baltimore, MD, USA

Alex D. Truong,
Division of Pulmonary and Critical Care Medicine, Johns Hopkins University, 1830 East Monument Street, 5th Floor, Baltimore, MD 21205, USA

Vinodh Bhoopathi,
Division of Dental Public Health, Boston University, Boston, MA, USA

Scott L. Zeger, and
Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA

Dale M. Needham
Division of Pulmonary and Critical Care Medicine, Johns Hopkins University, 1830 East Monument Street, 5th Floor, Baltimore, MD 21205, USA; Department of Physical Medicine and Rehabilitation, Johns Hopkins University, Baltimore, MD, USA

Eddy Fan: eddy.fan@jhmi.edu

Abstract

Objective—The goal of the paper is to determine inter-rater reliability of trained examiners performing standardized strength assessments using manual muscle testing (MMT).

Design, subjects, and setting—The authors report on 19 trainees undergoing quality assurance within a multi-site prospective cohort study.

Intervention—Inter-rater reliability for specially trained evaluators (“trainees”) and a reference rater, performing MMT using both simulated and actual patients recovering from critical illness was evaluated.

Measurements and results—Across 26 muscle groups tested by 19 trainee-reference rater pairs, the median (interquartile range) percent agreement and intraclass correlation coefficient (ICC; 95% CI) were: 96% (91, 98%) and 0.98 (0.95, 1.00), respectively. Across all 19 pairs, the ICC (95% CI) for the overall composite MMT score was 0.99 (0.98–1.00). When limited to actual patients, the ICC was 1.00 (95% CI 0.99–1.00). The agreement (kappa; 95% CI) in detecting clinically significant weakness was 0.88 (0.44–1.00).

Correspondence to: Eddy Fan, eddy.fan@jhmi.edu.

Electronic supplementary material The online version of this article (doi: 10.1007/s00134-010-1796-6) contains supplementary material, which is available to authorized users.

Conflicts of interest statement No potential conflicts of interest to disclose.

Conclusions—MMT has excellent inter-rater reliability in trained examiners and is a reliable method of comprehensively assessing muscle strength.

Keywords

Diagnostic techniques and procedures; Epidemiologic research design; Muscle strength; Muscle weakness; Physical examination; Reproducibility of results

Introduction

There is growing awareness of persistent neuromuscular complications after critical illness [1–5]. In studies evaluating interventions for improving muscle weakness, the reliable assessment of strength is key [6–8]. Standardized physical examination using manual muscle testing (MMT) is a widely accepted method for evaluating strength [9,10].

Prior studies evaluating the reproducibility of MMT have focused on a limited number of muscles or specific patient populations [11–14]. There is little data on the reliability of comprehensive MMT assessments in ICU survivors. Our objective is to determine the inter-rater reliability of trained examiners performing comprehensive MMT assessments using both actual patients recovering from critical illness and simulated patients.

Methods

In an ongoing study [15] approved by the Institutional Review Board of Johns Hopkins University, both clinicians and research assistants were trained to perform MMT with reproducibility evaluated against a single reference rater.

Manual muscle strength testing

MMT was scored using the 6-point Medical Research Council (MRC) scale [10]. Strength was evaluated bilaterally for 6 upper- and 7 lower-extremity muscle groups (total of 26 groups).

MRC scores for the 26 muscle groups were summed to yield a composite score of overall strength (range 0–130), as done previously [3,9,14]. Composite scores were also separately calculated for the upper (0–60) and lower extremities (0–70). Finally, an abbreviated composite score (0–60) was calculated based on a subset of 3 upper and 3 lower muscle groups, for comparison with a prior landmark study [9]. Composite scores were also dichotomized to designate patients with “clinically significant muscle weakness” if their score was <80% of the maximum score (i.e. average MRC score of <4 of 5 in all muscle groups) [3,4,9,14].

All personnel performing MMT completed multi-step training prior to their reliability assessments, including: review of a photo-illustrated MMT instruction manual; didactic teaching; and supervised practice by a trained staff member. The sole reference rater (NDC) was a physiotherapist with >30 years experience in both teaching and performing MMT across both clinical and research settings, particularly for ICU patients.

Nineteen different trainees underwent single-blinded MMT reliability evaluations with the reference rater. The trainees had various professional backgrounds (range of relevant experience with MMT): five physicians (1–10 years), four nurses (none), two respiratory therapists (none), five physiotherapists (6 months–5 years), one pharmacist (none), and two research assistants (none). Evaluations were conducted in a clinic setting using either an actual research participant (9 of 19) or a simulated patient (10 of 19) who effectively simulated a wide range of strength following training by the reference rater.

Statistical methods

For each of the 26 individual muscle groups, a median MMT score was separately calculated for the trainees versus reference rater and compared using the Wilcoxon signed rank test. For each trainee-reference rater pair, the overall composite score, upper and lower extremity composite scores, and the abbreviated composite score, were compared using percent agreement and intra-class correlation coefficients (ICC). Percent agreement was calculated by computing the number of individual muscle groups with exact agreement divided by the total number of muscles groups evaluated. To understand the direction of trainees' bias in disagreements among the pairs, we calculated a “proportion of lower disagreement” representing the proportion of all disagreements in which the trainee gave a lower score than reference rater. Using the nomenclature of Shrout and Fleiss [16], an ICC (2, 1) was calculated to provide a measure of inter-rater reliability that could be generalized beyond the study.

Reproducibility across all trainee-reference rater pairs was also determined using ICC for the overall and abbreviated composite scores. A kappa statistic was used to determine agreement in detecting the binary outcome of clinically significant muscle weakness. Statistical analyses were performed with Stata v.10.1 (College Station, TX, USA).

Results

Table 1 describes the range and median MMT score for each of the 26 muscle groups. Across these muscle groups, there were no statistically significant or clinically important differences in mean MMT score between the 19 pairs.

Table 2 describes the reproducibility of the composite scores for each of the 19 pairs. For all 26 muscles, the median (interquartile range) values for the percent agreement and ICC were: 96% (91, 98%) and 0.98 (0.95, 1.00). The median proportion of lower disagreement was 0% (0, 10%) indicating that trainees were biased toward higher strength scores versus reference rater. For the upper extremity, the median values for percent agreement and ICC were: 92% (87, 100%) and 0.99 (0.93, 1.00). The median proportion of lower disagreement was 0% (0, 10%). For the lower extremity, the median values for percent agreement and ICC were: 100% (93, 100%) and 1.00 (0.98, 1.00). The median proportion of lower disagreement was 0% (0, 0%).

For the median (interquartile range) overall composite MMT score, there was no clinically or statistically significant difference between trainees versus reference rater [96 (85–109) vs. 98 (83–107), $P = 0.052$]. Across all 19 trainee-reference rater pairs, the ICC (95% CI) for the overall, abbreviated, upper extremity, and lower extremity composite scores were: 0.99 (0.98–1.00), 0.99 (0.97–1.00), 0.97 (0.94–0.99), and 0.99 (0.98–1.00), respectively. When this analysis was restricted to the subgroup of assessments performed on the research participants, the ICC (95% CI) for the overall composite score was 1.00 (0.99–1.00), consistent with the overall analysis.

The kappa (95% CI) for agreement in detecting clinically significant weakness for all muscle groups, the upper extremities only, and the lower extremities only was 0.88 (0.44–1.00), 0.88 (0.44–1.00), and 1.00 (0.55–1.00), respectively. Analysis using the abbreviated composite score demonstrated a kappa (95% CI) of 1.00 (0.55–1.00).

Discussion

There is excellent reproducibility for MMT assessments across a wide range of specially trained research personnel who perform these evaluations. For the 26 muscle groups evaluated, MMT has substantial reproducibility with a median percent agreement and ICC of 96% and 0.98,

respectively. Furthermore, ICC (95% CI) for both the overall and abbreviated composite scores across all trainee-reference rater pairs was high at 0.99 (0.98–1.00) and 0.99 (0.97–1.00), respectively. These results were consistent when the analysis was limited to evaluations performed using only research participants (i.e. excluding simulated patients). Finally, there was substantial agreement between the trainees and reference rater for detecting clinically significant weakness using both overall and abbreviated composite scores.

Our study evaluates the reproducibility of a comprehensive MMT assessment in ICU survivors, who often experience long-term muscle weakness [1,2]. Our results are consistent with prior studies, evaluating other patient populations, that showed excellent percent agreement (range 82–98%) in composite MMT scores between raters [11,17]. Our findings were also similar to two studies of MMT in muscular dystrophy patients (ICC 0.75–0.96) [18,19]. Finally, our findings are consistent with a prior study [14] in Guillain-Barre patients, which reported an ICC of 0.96 for the abbreviated composite score. Our study is unique in evaluating the reproducibility of a comprehensive MMT examination, involving 26 muscle groups, within a study of ICU survivors. However, we are not specifically advocating for the evaluation of 26 muscle groups and the introduction of a new threshold (i.e. <104 out of 130), given the excellent inter-rater reliability demonstrated by the abbreviated composite score.

When MMT scores differed between trainees versus reference rater, the direction of bias was towards an overestimation of strength. This conservative bias indicates that trainees do not appear to overestimate weakness. Furthermore, despite variability in reliability across all trainee-reference rater pairs in composite scores, the ability to detect clinically significant weakness demonstrated substantial reliability.

The MMT training program in this study likely played an important role in these findings, consistent with prior studies which suggested that greater training and clinical experience improved MMT reliability [13,20].

ICU patients often experience prolonged immobility leading to substantial generalized weakness [2,5,9]. Therefore, a global measure of strength (e.g. composite MMT score) is important in gauging the severity of this weakness, and the impact of therapeutic interventions. In addition, the ability to ascertain muscle strength separately in upper and lower extremities may be helpful in prognosticating a patient's ability to perform specific functional tasks (e.g. ambulation, eating).

Our study has potential limitations. Approximately half of the study population were simulated patients. This study design may affect the generalizability of our findings; however, analysis of the subgroup of actual participants was consistent with the overall analysis. Furthermore, use of specially trained simulated patients allowed us to mimic a wide range of muscle weakness for comprehensive quality assurance testing versus use of real participants who may have less variability in strength across muscle groups and fatigue with repeated testing by both the trainee and reference rater. Finally, using an MRC score of <80% to define clinically significant weakness is arbitrary, as there has been limited research evaluating MMT scores and functional abilities. However, this methodology has been widely employed [3,4,9,14], making this threshold important for comparisons across studies.

In conclusion, MMT has excellent reproducibility when performed by both clinicians and research assistants who received specific training for research purposes. A composite score, combining MMT assessments from the upper and lower extremities, had high reproducibility for detecting clinically significant weakness. Among trained research staff, MMT is reliable for longitudinal and comprehensive assessment of strength.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research is supported by the National Institutes of Health (Acute Lung Injury SCCOR grant P050 HL 73994). EF is supported by a Fellowship Award from the Canadian Institutes of Health Research and a Detweiler Traveling Fellowship from the Royal College of Physicians and Surgeons of Canada. DMN is supported by a Clinician-Scientist Award from the Canadian Institutes of Health Research. The funding bodies had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

References

1. Herridge MS, Cheung AM, Tansey CM, Matte-Martyn A, Diaz-Granados N, Al-Saidi F, Cooper AB, Guest CB, Mazer CD, Mehta S, Stewart TE, Barr A, Cook D, Slutsky AS. Canadian Critical Care Trials Group. One-year outcomes in survivors of the acute respiratory distress syndrome. New Engl J Med 2003;348:683–693. [PubMed: 12594312]
2. Stevens RD, Dowdy DW, Micheals RK, Mendez-Tellez PA, Pronovost PJ, Needham DM. Neuromuscular dysfunction acquired in critical illness: a systematic review. Intensive Care Med 2007;33:1876–1891. [PubMed: 17639340]
3. Nanas S, Kritikos K, Angelopoulos E, Siafaka A, Tsikriki S, Poriazi M, Kanaloupiti D, Kontogeorgi M, Pratikaki M, Zervakis D, Routsi C, Roussos C. Predisposing factors for critical illness polyneuropathy in a multidisciplinary intensive care unit. Acta Neurol Scand 2008;118:175–181. [PubMed: 18355395]
4. Guarneri B, Bertolini G, Latronico N. Long-term outcome in patients with critical illness myopathy or neuropathy: the Italian multicenter CRIMYNE study. J Neurol Neurosurg Psychiatry 2008;79:838–841. [PubMed: 18339730]
5. Ali NA, O'Brien JM Jr, Hoffmann SP, Phillips G, Garland A, Finley JC, Almoosa K, Hejal R, Wolf KM, Lemeshow S, Connors AF Jr, Marsh CB. Midwest Critical Care Consortium. Acquired weakness, handgrip strength, and mortality in critically ill patients. Am J Respir Crit Care Med 2008;178:261–268. [PubMed: 18511703]
6. Hughes VA, Frontera WR, Wood M, Evans WJ, Dallal GE, Roubenoff R, Fiatarone Singh MA. Longitudinal muscle strength changes in older adults: influence of muscle mass, physical activity, and health. J Gerontol A Biol Sci Med Sci 2001;56:B209–B217. [PubMed: 11320101]
7. Steffensen BF, Lyager S, Werge B, Rahbek J, Mattsson E. Physical capacity in non-ambulatory people with Duchenne muscular dystrophy or spinal muscular atrophy: a longitudinal study. Dev Med Child Neurol 2002;44:623–632. [PubMed: 12227617]
8. Carin-Levy G, Greig C, Young A, Lewis S, Hannan J, Mead G. Longitudinal changes in muscle strength and mass after acute stroke. Cerebrovasc Dis 2006;21:201–207. [PubMed: 16401884]
9. De Jonghe B, Sharshar T, Lefaucheur JP, Authier FJ, Durand-Zaleski I, Boussarsar M, Cerf C, Renaud E, Mesrati F, Carlet J, Raphael JC, Outin H, Bastuji-Garin S. Groupe de Reflexion et d'Etude des Neuromyopathies en Reanimation. Paresis acquired in the intensive care unit: a prospective multicenter study. JAMA 2002;288:2859–2867. [PubMed: 12472328]
10. Medical Research Council/Guarantors of Brain. Aids to the examination of the peripheral nervous system. Bailliere Tindall; London: 1986.
11. Polard H, Lakay B, Tucker F, Watson E, Bablis P. Interexaminer reliability of the deltoid and psoas muscle test. J Manipulative Physiol Ther 2005;18:52–56.
12. Gregson JM, Leathley MJ, Moore P, Smith TL, Sharma AK, Watkins CL. Reliability of measure tones and muscle power in stroke patients. Age Aging 2000;29:223–228.
13. Florence JM, Pandya S, King WM, Robison JD, Signore LC, Wentzell M, Province MA. Clinical trials in Duchenne dystrophy: standardization and evaluation of reliability procedures. Phys Ther 1984;64:41–45. [PubMed: 6361809]

14. Kleyweg RP, Meche FGA, Schmitz PI. Interobserver agreement in the assessment of muscle strength and functional abilities Guillain-Barre syndrome. Muscle Nerve 1991;14:1103–1109. [PubMed: 1745285]

15. Needham DM, Dennison CR, Dowdy DW, Mendez-Tellez PA, Ciesla N, Desai SV, Sevransky J, Shanholtz C, Scharfstein D, Herridge MS, Pronovost PJ. Study protocol: the improving care of acute lung injury patients (ICAP) study. Crit Care 2005;10:R9. [PubMed: 16420652]

16. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–427. [PubMed: 18839484]

17. Perry J, Weiss WB, Burnfield JM, Gronley JK. The supine hip extensor manual muscle test: a reliability and validity study. Arch Phys Med Rehabil 2004;85:1345–1350. [PubMed: 15295763]

18. Barr AE, Diamond BE, Wade CK, Harashima T, Pecorella WA, Potts CC, Rosenthal H, Fleiss JL, McMahon DJ. Reliability of testing measures in Duchenne or Becker muscular dystrophy. Arch Phys Med Rehabil 1991;72:315–319. [PubMed: 2009048]

19. Escolar DM, Henricson EK, Mayhew J, Florence J, Leshner R, Patel KM, Clemens PR. Clinical evaluator reliability for quantitative and manual muscle testing measures of strength in children. Muscle Nerve 2001;24:787–793. [PubMed: 11360262]

20. Caruso W, Leisman G. A force/displacement analysis of muscle testing. Percept Mot Skills 2000;91:683–692. [PubMed: 11065332]

Table 1

Manual muscle test score summary statistics for individual muscle groups

Muscle group	Gold standard score				Trainee Score				*P* value[b]
	Number assessed[a]	Minimum	Maximum	Median (IQR)	Number assessed[a]	Minimum	Maximum	Median (IQR)	
Shoulder abduction right	19	1	5	4.0 (2.0–5.0)	19	1	5	4.0 (2.0–5.0)	0.16
Shoulder abduction left	19	2	5	4.0 (4.0–5.0)	19	2	5	4.0 (4.0–5.0)	0.16
Shoulder flexion right	19	2	5	4.0 (2.0–5.0)	19	2	5	4.0 (2.0–5.0)	1.00
Shoulder flexion left	19	2	5	4.0 (4.0–5.0)	19	2	5	4.0 (4.0–5.0)	1.00
Elbow flexion right	19	2	5	4.0 (3.0–5.0)	19	2	5	4.0 (3.0–5.0)	0.32
Elbow flexion left	18	1	5	5.0 (4.0–5.0)	18	1	5	5.0 (4.0–5.0)	0.32
Elbow extension right	14	2	5	5.0 (3.0–5.0)	15	2	5	5.0 (3.0–5.0)	0.32
Elbow extension left	19	0	5	5.0 (2.0–5.0)	19	0	5	5.0 (2.0–5.0)	0.32
Wrist flexion right	19	1	5	5.0 (3.0–5.0)	19	1	5	5.0 (4.0–5.0)	0.30
Wrist flexion left	18	1	5	4.5 (3.0–5.0)	19	1	5	5.0 (2.0–5.0)	0.32
Wrist extension right	19	1	5	4.0 (3.0–5.0)	19	1	5	4.0 (3.0–5.0)	0.32
Wrist extension left	19	1	5	4.0 (4.0–5.0)	19	1	5	4.0 (1.0–5.0)	0.16
Hip flexion right	19	1	5	4.0 (3.0–5.0)	19	1	5	4.0 (3.0–5.0)	0.32
Hip flexion left	19	1	5	5.0 (3.0–5.0)	19	1	5	5.0 (3.0–5.0)	1.00
Hip extension right	17	2	5	4.0 (2.0–5.0)	17	2	5	4.0 (2.0–5.0)	0.32
Hip extension left	17	2	5	5.0 (4.0–5.0)	17	2	5	5.0 (4.0–5.0)	1.00
Hip abduction right	19	2	5	4.0 (2.0–5.0)	18	2	5	4.0 (3.0–5.0)	1.00
Hip abduction left	17	1	5	4.0 (3.0–5.0)	17	1	5	4.0 (3.0–5.0)	1.00
Knee flexion right	19	1	5	4.0 (3.0–5.0)	19	1	5	4.0 (3.0–5.0)	0.32
Knee flexion left	19	1	5	4.0 (3.0–5.0)	19	1	5	4.0 (3.0–5.0)	1.00
Knee extension right	19	2	5	5.0 (3.0–5.0)	19	2	5	5.0 (3.0–5.0)	1.00
Knee extension left	19	2	5	5.0 (3.0–5.0)	18	2	5	5.0 (3.0–5.0)	1.00
Ankle dorsiflexion right	19	0	5	5.0 (4.0–5.0)	19	0	5	5.0 (4.0–5.0)	1.00
Ankle dorsiflexion left	19	1	5	5.0 (2.0–5.0)	19	1	5	5.0 (2.0–5.0)	1.00
Ankle plantarflexion right	15	1	5	4.0 (2.0–5.0)	15	1	5	4.0 (2.0–5.0)	0.16
Ankle plantarflexion left	18	1	5	3.5 (2.0–4.0)	17	1	5	4.0 (3.0–5.0)	0.32

IQR interquartile range

[a] Not all muscles could be tested in every patient

[b] Wilcoxon signed rank test for equality of median scores

Table 2

Inter-rater reliability for individual MMT scores across 19 evaluations

Examiner pair	Patient type	All muscles (*n* = 26)				Upper extremity muscles (*n* = 12)				Lower extremity muscles (*n* = 14)			
		Number Assessed[a]	Percent agreement (%)	Disagreement proportion[b] (%)	ICC	Number assessed[a]	Percent agreement (%)	Disagreement proportion[b] (%)	ICC	Number assessed[a]	Percent agreement (%)	Disagreement proportion[b] (%)	ICC
1	S	26	92	0	0.97	12	100	0	1.00	14	86	0	0.82
2	S	24	79	20	0.69	12	58	20	0.62	12	100	0	1.00
3	S	22	86	100	0.96	11	82	100	0.97	11	91	0	0.66
4	S	20	90	100	0.92	11	82	100	0.91	9	100	0	1.00
5	S	26	96	0	0.91	12	92	0	0.82	14	100	0	1.00
6	S	24	96	0	0.96	11	100	0	1.00	13	92	0	0.94
7	S	24	92	0	0.98	11	100	0	1.00	13	85	0	0.96
8	S	26	96	0	0.99	12	92	0	0.98	14	100	0	1.00
9	S	23	96	0	0.99	10	90	0	0.98	13	100	0	1.00
10	S	26	100	0	1.00	12	100	0	1.00	14	100	0	1.00
11	P	26	85	0	0.86	12	67	0	0.45	14	100	0	1.00
12	P	26	100	0	1.00	12	100	0	1.00	14	100	0	1.00
13	P	26	92	0	0.98	12	92	0	0.99	14	93	0	0.98
14	P	26	88	33	0.96	12	83	50	0.94	14	93	0	0.98
15	P	25	96	100	0.99	11	91	100	0.99	14	100	0	1.00
16	P	26	100	0	1.00	12	100	0	1.00	14	100	0	1.00
17	P	25	96	0	0.94	12	92	0	0.85	13	100	0	1.00
18	P	26	100	0	1.00	12	100	0	1.00	14	100	0	1.00
19	P	26	100	0	1.00	12	100	0	1.00	14	100	0	1.00
Median		26	96	0	0.98	12	92	0	0.99	14	100	0	1.00

ICC intraclass correlation coefficient, *S* simulated participant, *P* research participant

[a] Not all muscles could be tested

[b] Proportion of lower disagreement when trainee score is less than the gold standard scores