Assessment of Limb Muscle Strength in Critically Ill Patients: A Systematic Review

Goele Vanpee, MSc, PT[1,2]; Greet Hermans, PhD, MD[3]; Johan Segers, MSc, PT[1,2]; Rik Gosselink, PhD, PT[1,2]

Objectives: To determine the reliability of volitional and nonvolitional limb muscle strength assessment in critically ill patients and to provide guidelines for the implementation of limb muscle strength assessment this population.

Data Sources: The following computerized bibliographic databases were searched with MeSH terms and keywords or combinations: MEDLINE through PubMed and Embase through Embase.com.

Study Selection: Articles were screened by two independent reviewers. Included studies were all performed in humans and were original articles. The research population exists of adult, critically ill patients or ICU survivors of either sex, and those admitted to a medical, surgical, respiratory, or mixed ICU. A study was included if reliability of muscle strength measurements was determined in this population.

Data Extraction: Data on baseline characteristics (country, study population, eligibility, age, setting and method, and equipment of limb muscle strength assessment) and reliability scores were obtained by two independent reviewers.

Data Synthesis: Data of six observational studies were analyzed. Interrater reliability of the Medical Research Council scale for individual muscle groups varied from "fair" or "substantial" (weighted κ, 0.23–0.64) to "very good" agreement (weighted κ, 0.80–0.96). Interrater reliability of the Medical Research Council-sum score was found to be very good in all four studies (intraclass correlation coefficients, 0.86–0.99 or Pearson product moment correlation coefficient = 0.96). Interrater reliability of handheld dynamometry was comparable between two studies (intraclass correlation coefficients, 0.62–0.96). Interrater reliability of handgrip dynamometry was very good in two studies (intraclass correlation coefficients, 0.89–0.97). Intrarater reliability of handheld dynamometry and handgrip dynamometry was assessed in one study, and results were very good (intraclass correlation coefficients > 0.81). No studies were obtained on reliability of nonvolitional muscle strength assessment.

Conclusions: Voluntary muscle strength measurement has proven reliable in critically ill patients provided that strict guidelines on adequacy and standardized test procedures and positions are followed. (*Crit Care Med* 2014; 42:701–711)

Key Words: adult; critical illness; muscle strength; muscle strength dynamometer; muscle weakness; reliability of results

The development of limb muscle weakness is a clinically important feature in critically ill patients treated in the ICU. It is associated with a prolonged duration of mechanical ventilation and a protracted ICU stay. Furthermore, muscle weakness is associated with an increased risk for morbidity and mortality (1, 2) and has devastating effects on functional outcome and social well-being even years after the ICU stay (3, 4). Evaluation of muscle strength is important to detect the presence of muscle weakness (5), to make the appropriate selection of exercise modalities to counteract the development of muscle weakness, and to evaluate the effect of clinical interventions. Therefore, objective, reliable, and sensitive tools to measure muscle strength are necessary.

Volitional and nonvolitional muscle strength assessment tools are available. The Medical Research Council (MRC) scale is a volitional muscle strength test for cooperative patients. The MRC scale is a categorical scale to measure the entire range of muscle strength, from 0 (no visible or palpable muscle contraction) through 5 (movement through the complete range of motion against gravity and maximum resistance) (6). It produces ordinal data and is partly a subjective evaluation of muscle strength. Grades up to 3 may provide an objective score for strength assessment in patients with more profound weakness. However, several studies indicate difficulties in differentiation between grades 4 and 5, inaccuracy in identifying muscle weakness in patients compared with healthy subjects, and a

[1]Department Rehabilitation Sciences, Faculty of Kinesiology and Rehabilitation Sciences, KU Leuven, Leuven, Belgium.

[2]Department of Intensive Care, University Hospitals Leuven, Leuven, Belgium.

[3]Medical Intensive Care Unit of the Department of General Internal Medicine, University Hospitals Leuven, Leuven, Belgium.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccmjournal).

The authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: rik.gosselink@faber.kuleuven.be

Copyright © 2013 by the Society of Critical Care Medicine and Lippincott Williams & Wilkins

DOI: 10.1097/CCM.0000000000000030

lack of sensitivity to detect progress in muscle strength when applied to stronger muscle groups (grades 4 and 5) (6–8).

Kleyweg et al (9) developed the MRC-sum score to identify general peripheral muscle weakness. This sum score comprehends the individual score for (bilateral) muscle groups of the upper limbs and the lower limbs. The MRC and MRC-sum score have been implemented in the examination of muscle strength in critically ill patients to assess ICU-acquired weakness (ICUAW) (2, 10).

Handheld dynamometry (HHD) and handgrip dynamometry (HGD) have been designed to measure volitional isometric muscle strength more objectively in patients who are cooperative and have a score of 3 or more on the MRC (11, 12). HHD and HGD express muscle strength in continuous data. Two methods of isometric testing with HHD have been described. The make technique requires the patient to exert a maximal isometric contraction while the examiner holds the dynamometer in a fixed position. The break technique, in contrast, requires the examiner to overpower a maximal effort by the patient, thereby producing a measurement of eccentric muscle strength (13).

Volitional strength measurements may be affected by several factors, such as awareness, cooperation, and motivation of the patients or the applied test procedures and positions (14). Specifically in critically ill patients, these factors may hamper feasibility, sensitivity, and reliability of the measurements. Nonvolitional muscle strength measurements, such as electrical and magnetic neuromuscular twitch stimulation (15), can be applied earlier in the recovery process, since these measurements are independent of the adequacy and motivation of the patient.

The first objective of this systematic review is to provide an overview on the reliability of different methods available to assess limb muscle strength in critically ill patients. The second objective is to define recommendations for the assessment of limb muscle strength in critically ill patients.

METHODS

Data Sources and Searches

Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for systematic reviews were followed. In the campus library Biomedical sciences of the Katholieke Universiteit Leuven (KU Leuven), the following computerized bibliographic databases were searched with MeSH terms and keywords or combinations: MEDLINE through PubMed (September 1979–October 2012) and Embase through Embase.com (1992–October 2012). The full search strategy is included in **Table 1**. October 2012 was the last run of the search string. Reference lists were screened to identify additional relevant studies.

Study Selection

Studies included in the systematic review were all performed in humans and were original articles (e.g., not an abstract, review, or editorial). The research population exists of adult (> 18 yr), critically ill patients of either sex, and those admitted to a medical, surgical, respiratory, or mixed ICU. ICU survivors are also part of the research population. A study was included if reliability of muscle strength measurements was performed in critically ill patients or ICU survivors. Reliability is defined as the extent to which results of muscle strength assessment are the same for repeated measurements under several conditions: for example, by different persons on the same occasion (intertester) or by the same person on different occasions (intratester). Two reviewers (G.V., J.S.) independently checked the titles, abstracts, and reference lists of studies retrieved by the literature search. The full texts of all potentially relevant studies were obtained, and the reviewers decided again independently which trials fitted the inclusion criteria. In case of disagreement between the two reviewers, there was discussion to reach consensus. If necessary, a third reviewer (R.G.) made the decision regarding inclusion of the article.

Data Extraction

Data on baseline characteristics (country where the study was performed, study population, eligibility, equipment, age, setting, and the method that was used for the assessment of limb muscle strength) and reliability scores were obtained. Two reviewers (G.V., J.S.) independently extracted data. In case of disagreement between the two reviewers (G.V., J.S.), there was discussion to reach consensus. If necessary, a third reviewer (R.G.) made the decision. In case of missing data, we contacted the authors.

Quality Assessment

Assessment of methodological quality was performed using the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist with 4-point scale (14). The COSMIN checklist is recommended for use in systematic reviews of measurement properties in observational studies. The COSMIN checklist consists of nine boxes (one box for each measurement property). For this systematic review, only the methodological quality of the measurement property "reliability" of a study was evaluated by two independent reviewers (G.V., J.S.). Since it has been reported that interrater reliability of the COSMIN checklist may be influenced by interpretation on different items of the checklist (16), rules were set in advance on how to score different items. Agreement for quality assessment between reviewers was calculated with a κ score by an independent third person. Whenever disagreement between the two reviewers (G.V., J.S.) occurred after scoring all the included articles for quality, there was discussion to reach consensus. If necessary, a third reviewer (R.G.) made the decision. Quality assessment scores are listed in **Table 2**.

Data Analysis and Synthesis

An overview of baseline characteristics and reliability data from original articles is listed in **Table 3**. For quality assessment of the original articles, SPSS (SPSS Inc, Chicago, IL) was used to calculate the interobserver agreement on quality scoring with the COSMIN checklist. An unweighted κ score between observers (G.V., J.S.) was calculated by a third independent person. κ scores were interpreted according to Landis and Kock (17).

TABLE 1. Search Strategy for Reliability of Muscle Strength Assessment in Critically Ill Patients

Medline through PubMed	(“Critical Illness”[Mesh] or “critical ill”[tiab] or “critical illness”[tiab] or “critically ill”[tiab] or “Intensive Care Units”[Mesh] or Intensive Care Unit*[tiab] or critical care unit*[tiab] or ICU[tiab]) AND (“Muscle Strength Dynamometer”[Mesh] or Dynamomet* or ((“Muscle Strength”[Mesh] or “Muscle Weakness”[Mesh] or muscle* or muscular or strength) and (assessment* or measur* or screen* or testing or test or tests)) or ((muscle or muscular or handgrip) and strength) or “Isometric Contraction”[Mesh] or ((isometric or muscular) and contraction*) or (MRC or “manual muscle strength” or manual muscle test* or MMT or “manual strength” or “medical research council scale”) or ((non-invasive or noninvasive or “non invasive”) and (muscle force assessment* or muscle function assessment*))) AND (“Reproducibility of Results”[Mesh] or “Predictive Value of Tests”[Mesh] or reliabilit* or reliable or validat* or validity or ((interobserver or “inter observer” or “inter-observer” or intraobserver or “intra observer” or “intra-observer”) and agreement*) or ((inter or intra) and (rater* or tester or tester or testing)) or intertest* or intratest* or interrater* or intrarater*) NOT (animals[mh] not humans[mh])
Embase through Embase.com	**#1** “critical illness”/exp **#2** “critically ill patient”/exp **#3** “intensive care unit”/exp **#4** “critical ill”:ab,ti or “critical illness”:ab,ti or “critically ill”:ab,ti and [embase]/lim **#5** “intensive care unit”:ab,ti or “intensive care units”:ab,ti or “critical care unit”:ab,ti or “critical care units”:ab,ti or icu:ab,ti and [embase]/lim **#6** #1 or #2 or #3 or #4 or #5 **#7** “dynamometer”/exp **#8** dynamomet* and [embase]/lim **#9** “muscle strength”/exp **#10** “muscle weakness”/de **#11** assessment* or measur* or screen* or testing or test or tests and [embase]/lim **#12** #9 or #10 **#13** #11 and #12 **#14** (muscle* or muscular or strength) NEAR/1 (assessment* or measur* or screen* or testing or test or tests) and [embase]/lim **#15** (muscle or muscular or handgrip) NEAR/1 strength and [embase]/lim **#16** “muscle isometric contraction”/exp **#17** (isometric or muscular) NEAR/1 contraction* and [embase]/lim **#18** mrc or “manual muscle strength” or “manual muscle test” or “manual muscle tests” or “manual muscle testing” or mmt or “manual strength” or “medical research council scale” and [embase]/lim **#19** (“non invasive” or “non-invasive” or noninvasive) NEAR/1 (“muscle force assessment” or “muscle function assessment”) and [embase]/lim **#20** #7 or #8 or #13 or #14 or #15 or #16 or #17 or #18 or #19 **#21** “reproducibility”/exp **#22** “predictive value”/exp **#23** reliabilit* or reliable or validat* or “validity”/exp or validity and [embase]/lim **#24** (interobserver or “inter observer” or “inter-observer” or intraobserver or “intra observer” or “intra-observer”) NEAR/1 agreement* and [embase]/lim **#25** (inter or intra) NEAR/1 (rater* or tester or tester or testing) and [embase]/lim **#26** intertest* or intratest* or interrater* or intrarater* and [embase]/lim **#27** #21 or #22 or #23 or #24 or #25 or #26 **#28** “test retest reliability”/exp

(*Continued*)

TABLE 1. (*Continued*). Search Strategy for Reliability of Muscle Strength Assessment in Critically Ill Patients

#29	#27 or #28
#30	"humans" or "humans"/exp or humans and [embase]/lim
#31	"animal" or "animal"/exp or animal and [embase]/lim
#32	#31 not #30
#33	#6 and #20 and #29
#34	#33 not #32

RESULTS

Description of Studies

The search strategy resulted in 196 publications of which 160 were identified by PubMed and 36 results were identified by Embase (**Fig. 1**). After removal of duplicates ($n = 15$), 181 publications were screened on title and abstract. A total of 159 publications were excluded for reasons described in Figure 1. The remaining 22 publications were screened on full text. Sixteen publications were excluded for reasons described in Figure 1. Reference checking did not reveal additional included articles. Finally, six articles were eligible for this review (Fig. 1; Table 3). No reliability studies were obtained for nonvolitional muscle strength assessment.

Methodological Quality

Methodological quality of all studies was screened on the measurement property reliability (Table 2). Agreement on assignment of methodological quality between both authors (G.V., J.S.) was "very good" (17) with a κ score of 0.86.

Reliability of the MRC and MRC-sum score was reported in two studies (18, 19), and two studies included reliability of the MRC-sum score (2, 20). Some of these four studies had lower scores on item 2, item 3, and item 7. Item 2 was scored as fair in two studies (2, 20). The other two studies (18, 19) scored excellent on this item. Three of the four studies (2, 19, 20) scored only "fair" or "poor" on item 3 "adequacy of sample size." Item 7 "stability of patients between measurements" was scored "fair" in two of four studies (2, 20).

Reliability of HHD and HGD was investigated in three different studies (18, 21, 22). In all three studies (18, 21, 22), item 2 was scored as "not applicable." Item 3 was scored fair or poor in two studies (18, 22). Only one of the three studies scored fair on item 7.

In all six studies, statistical analysis was scored good or excellent in item 11, 12, 13, and 14 of the COSMIN checklist. Only one study scored fair on item 11 (2).

TABLE 2. Quality Assessment of the Included Studies

Box B: Reliability	Ali et al (2)	Hough et al (19)
Design Requirements	**MRC Sum**	**MRC and MRC Sum**
1. Was the percentage/number of missing items given?	Excellent	Excellent
2. Was there a description of how missing data were handled?	Fair	Excellent
3. Was the sample size included in the analysis adequate?	Poor	Fair
4. Were there at least two measurements available?	Excellent	Excellent
5. Were the administrations independent?	Excellent	Excellent
6. Was the time interval stated?	Excellent	Excellent
7. Were patients stable in the interim period on the construct to be measured?	Fair	Good
8. Was the time interval appropriate?	Excellent	Fair
9. Were test conditions similar for both measurements?	Excellent	Poor
10. Were there any important flaws in the design or methods of the study?	Excellent	Excellent
Statistical methods		
11. For continuous scores: was an intraclass correlation coefficient calculated?	Fair	Excellent
12. For dichotomous/nominal/ordinal scores: was κ calculated?	Not applicable	Excellent
13. For ordinal scores: was a weighted κ calculated?	Not applicable	Excellent
14. For ordinal scores: was the weighting scheme described?	Not applicable	Excellent

MRC = medical research council, HHD = Handheld dynamometry.

MRC

Interrater reliability of the MRC score for individual muscle groups was calculated in two studies (18, 19). In total, 105 participants were included of which 85 were critically ill patients. Hough et al (19) revealed weighted κ coefficients (0.23–0.64), indicating fair to substantial reliability scores (17). Hermans et al (18) found very good agreement (weighted κ, 0.80–0.96) (17).

Interrater reliability for the MRC-sum score was calculated in four studies (2, 18–20) and included a total of 136 participants of which 97 participants were critically ill patients treated in a medical or surgical ICU. Intraclass correlation coefficients (ICC) of three studies (18–20) varied from 0.83 to 0.99, and Pearson product moment correlation coefficient equal to 0.96 was found in one study (2). Extrapolation of MRC score occurred in 57 of 900 measured muscle groups in the study by Hermans et al (18) and in 29 of 360 measured muscle groups in the study by Hough et al (19). No studies reported on intrarater reliability of the MRC.

Dynamometry

Interrater reliability of isometric muscle strength with HHD (using the make method) was investigated in two studies (21, 22). In total, 56 critically ill patients participated in these studies. ICC scores calculated by Vanpee et al (21) varied from 0.76 to 0.96. Baldwin et al (22) observed somewhat lower scores, especially for left elbow flexion with ICC (0.62–0.92). Intrarater reliability of isometric muscle strength with HHD was assessed in 17 critically ill patients (22). Intrarater reliability scores were slightly higher than interrater reliability scores for all muscle groups (ICC 0.82–0.91), with exception of left elbow flexion (ICC 0.42) (22).

Interrater reliability of isometric handgrip strength was assessed in two studies with in total 63 critically ill patients participating. ICC varied from 0.93 to 0.97 (18) and from 0.89 to 0.92 (22). Intrarater reliability of HGD was assessed in 17 critically ill patients and ICC varied from 0.85 to 0.91 (22).

DISCUSSION

Interrater reliability of MRC score for individual muscle groups (18, 19) varied from fair or substantial to very good agreement (17), whereas the reliability for the MRC-sum score was very good in all studies (2, 18–20). Interrater reliability of HHD was good to very good (21, 22), whereas interrater reliability of HGD was very good (18, 22). Intrarater reliability of HHD and HGD was very good (22).

Methodological Quality

Methodological quality of the studies was good or excellent for the majority of the items. In general, methodological quality of items 2, 3, and 7 of the COSMIN checklist was compromised. Item 2 implies that attention should be paid to clarify on how missing data are handled, and item 3 indicates that sample sizes should be adequate. However, it should be noticed that the COSMIN checklist was mainly developed to investigate the methodological quality of measurement properties of health status questionnaires, requiring large sample sizes between 30 and 100 subjects (23). Baldwin et al (22) included a power calculation and reported that 12 patients were required to achieve

Fan et al (20)	Vanpee et al (21)	Hermans et al (18)		Baldwin et al (22)	
MRC Sum	HHD	MRC and MRC Sum	Handgrip Force	HHD	Handgrip Force
Excellent	Excellent	Excellent	Good	Excellent	Excellent
Fair	Not applicable	Excellent	Not applicable	Not applicable	Not applicable
Poor	Good	Good	Fair	Poor	Poor
Excellent	Excellent	Excellent	Excellent	Excellent	Excellent
Good	Excellent	Excellent	Excellent	Excellent	Excellent
Fair	Excellent	Excellent	Excellent	Excellent	Excellent
Fair	Good	Good	Good	Fair	Fair
Not applicable	Excellent	Excellent	Excellent	Excellent	Excellent
Excellent	Excellent	Excellent	Excellent	Excellent	Excellent
Poor	Excellent	Excellent	Excellent	Excellent	Excellent
Excellent	Good	Excellent	Excellent	Excellent	Excellent
Not applicable	Not applicable	Excellent	Not applicable	Not applicable	Not applicable
Not applicable	Not applicable	Excellent	Not applicable	Not applicable	Not applicable
Not applicable	Not applicable	Good	Not applicable	Not applicable	Not applicable

TABLE 3. Characteristics and Results of the Included Studies

Hough et al (19)	
Country	United States
Population	10 critically ill patients, 20 ICU survivors admitted to the ward
Eligibility	Mechanical ventilation ≥ 3 d
Age	Mean: 49, SD 15
Measuring method	MRC scale (bilateral, 12 muscle groups were measured), MRC-sum score
Equipment	None
Setting	1 physician, 1 medical resident
Reliability	Interrater reliability
Reliability scores	MRC of individual muscle groups:
	Shoulder abduction: right, weighted κ = 0.51; left, weighted κ = 0.36
	Elbow flexion: right, weighted κ = 0.35; left, weighted κ = 0.23
	Wrist extension: right, weighted κ = 0.56; left, weighted κ = 0.44
	Hip flexion: right, weighted κ = 0.47; left, weighted κ = 0.32
	Knee extension: right, weighted κ = 0.29; left, weighted κ = 0.29
	Ankle dorsiflexion: right, weighted κ = 0.64; left, weighted κ = 0.32
	MRC-sum score: ICC = 0.83
Hermans et al (18)	
Country	Belgium
Population	75 critically ill patients (database 1), 46 critically ill patients (database 2)
Eligibility	Admitted to the ICU ≥ 7 d
Age	Median, 59; IQR, 52–71 (database 1)
	Median, 54; IQR, 47–68 (database 2)
Measuring method	MRC scale (bilateral, 12 muscle groups were measured) (database 1)
	MRC-sum score
	Handgrip strength (database 2)
Equipment	Hydraulic handgrip dynamometer (Jamar Preston, Jackson, MI)
Setting	Two physiotherapists (database 1)
	Two physiotherapists (database 2)
Reliability	Interrater reliability
Reliability scores	MRC individual muscle groups (database 1)
	Upper limbs muscle groups: weighted κ = 0.80, ICC = 0.92
	Lower limb muscle groups: weighted κ = 0.86, ICC = 0.96
	Proximal muscle groups: weighted κ = 0.84, ICC = 0.93
	Middle muscle groups: weighted κ = 0.80, ICC = 0.88
	Distal muscle groups: weighted κ = 0.83, ICC = 0.95
	MRC-sum score
	ICC = 0.95, weighted κ = 0.83
	Handgrip strength (database 2)
	Right, ICC = 0.93; left, ICC = 0.97

(*Continued*)

TABLE 3 (*Continued*). Characteristics and Results of the Included Studies

Fan et al (20)	
Country	United States
Population	Nine patients recovering from critical illness and 10 simulated patients
Eligibility	ND
Age	ND
Measuring method	MRC-sum score (bilateral, 26 muscle groups were measured)
Equipment	None
Setting	Physician, nurse, physiotherapist, respiratory therapist, pharmacist, research assistant
Reliability	Interrater reliability
	Detection of clinically significant weakness
Reliability scores	MRC sum
	ICC = 0.99
	MRC sum (12 muscle groups) for comparison with other studies
	ICC = 0.99
	MRC sum (upper extremities)
	ICC = 0.97
	MRC sum (lower extremities)
	ICC = 0.99
Ali et al (2)	
Country	United States
Population	12 critically ill patients
Eligibility	Mechanical ventilation ≥ 5 d
Age	Mean, 57.7; SD, 15.5
Measuring method	MRC sum (bilateral, 12 muscle groups were measured)
Equipment	Hydraulic handgrip dynamometer (Jamar; Sammons Preston Rolyan, Bolingbrook, IL)
Setting	Two physicians
Reliability	Interrater reliability
Reliability scores	Pearson correlation coefficient = 0.96
Vanpee et al (21)	
Country	Belgium
Population	39 critically ill patients, 51 test-retest sessions
Eligibility	Admitted to the ICU ≥ 7 d
Age	Median, 64; IQR, 53–72
Measuring method	Handheld dynamometry (unilateral, six muscle groups were measured)
Equipment	Handheld electronic dynamometer (CompuFet 2; Biometrics, Almere, the Netherlands)
Setting	Two physiotherapists
Reliability	Interrater reliability
Reliability scores	Shoulder abduction: ICC = 0.91
	Elbow flexion: ICC = 0.94
	Wrist extension: ICC = 0.96

(*Continued*)

TABLE 3 (*Continued*). Characteristics and Results of the Included Studies

	Hip flexion: ICC = 0.80
	Knee extension: ICC = 0.94
	Ankle dorsiflexion: ICC = 0.76
Baldwin et al (22)	
Country	Australia
Population	17 critically ill patients
Eligibility	Admitted to the ICU ≥ 5 d
Age	Median, 78; IQR, 46–82
Measuring method	Handgrip strength (bilateral)
	Handheld dynamometry (bilateral, four muscle groups were measured)
Equipment	Hydraulic handgrip dynamometer (Jamar; Sammons Preston Rolyan)
	Handheld electronic dynamometer (model 01163; Lafayette Instrument, Lafayette, IN)
Setting	Two physicians
Reliability	Intrarater reliability
	Interrater reliability
Reliability scores	Intrarater reliability handgrip strength
	Right, ICC = 0.91; left, ICC = 0.85
	Intrarater reliability elbow flexion
	Right, ICC = 0.82; left, ICC = 0.42
	Intrarater reliability knee extension
	Right, ICC =0.90; left, ICC = 0.91
	Interrater reliability handgrip strength
	Right, ICC = 0.92; left, ICC = 0.89
	Interrater reliability elbow flexion
	Right, ICC = 0.71; left, ICC = 0.62
	Interrater reliability knee extension
	Right, ICC = 0.84; left, ICC = 0.79

MRC = medical research council, ICC = intraclass correlation coefficient, IQR = interquartile range, ND = not defined.

80% power to assess reliability of dynamometry. This suggests that sample sizes are not required to be as large as suggested by the COSMIN checklist to be sufficient for the reliability analyses presented in the included studies. Item 7 implies in the context of the present study that attention should be paid to hemodynamic and neurological (cooperation) stability at test and retest moments.

The good results on items 11, 12, 13, and 14 indicate that the correct statistical methods for analysis of results in the included studies were performed.

MRC Scale

Assessors gained clinical experience with manual muscle testing in ICU patients before the start of the study (18, 19) to improve reliability of muscle strength. The differences in interrater reliability for individual muscle groups between the two studies can be explained by differences in level of cooperation requested in these studies. Hough et al (19) requested a correct response on three of the five standardized questions for adequacy as an inclusion criterion, whereas Hermans et al (18) included only fully adequate patients with a score of 5 on the five standardized questions.

Another potentially confounding factor explaining differences in reliability is the use of standardized test positions. Hough et al (19) positioned their patients in either the sitting or supine position at test and retest, depending on the patient's clinical condition, whereas Hermans et al (18) tested all patients in supine position.

Differences in patient characteristics may also account for differences in reliability. Hough et al (19) performed the first strength evaluation much earlier during the ICU admission (median, 8 d; interquartile range [IQR], 6–12) and obtained

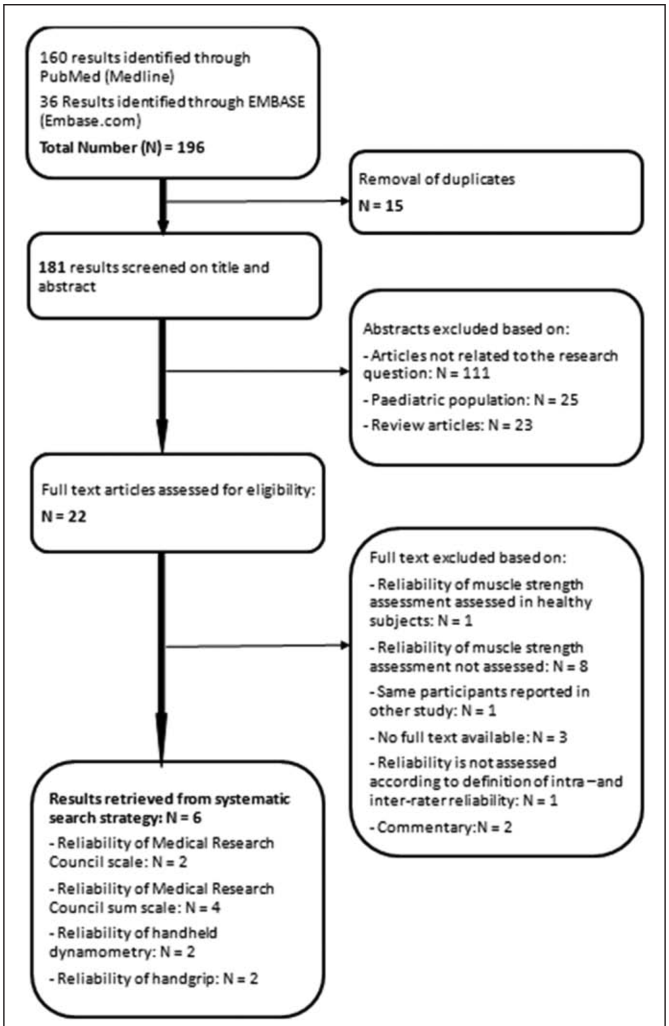

Figure 1. Search strategy flowchart.

higher MRC-sum scores (median, 55; IQR, 49–58) compared with Hermans et al (18) (median MRC, 48; IQR, 46–54; median first evaluation at 22 d; IQR, 15–30). Other studies did not report length of ICU stay prior to the first muscle strength assessment (2, 20). No learning effects or fatigue were induced by repeated measurements within a short time interval (18). The procedure for handling of missing values due to peripheral or central nervous lesions, orthopedic conditions, amputation, or others was reported in two studies. Extrapolation of the value from the identical contralateral muscle group or the value of the ipsilateral group of muscles in the same proximity was used to calculate the MRC-sum score (18, 19).

Interobserver agreement on MRC-sum scores was mostly evaluated by calculating ICCs according to Shrout and Fleiss (24) and indicated very good agreement (ICCs > 0.83) (18–20). However, Kleyweg et al (9) suggested that a difference of less than 10% in MRC-sum score (i.e., a difference of less than 6 points) indicates good agreement. In the study by Hough et al (19), interobserver agreement on the MRC-sum score as a continuous outcome differed by 10% or more in 23% of the patients.

Dynamometry

Interrater reliability of HHD was good to excellent in the two studies (21, 22). Only for elbow flexion and knee extension, interrater reliability scores were less strong in the study by Baldwin et al (22) (ICC = 0.71 and 0.84) compared with Vanpee et al (21) (ICC = 0.94 and 0.94). Several differences in methodology may explain these results. First, the level of cooperation was assessed differently. Vanpee et al (21) required patients to be fully awake and adequate (score of 5 on the five standardized questions), whereas Baldwin et al (22) required a minimum score of more than 8 of 10 on the Attention Screening Examination. However, only 13 of 17 patients obtained this score. Second, in both studies, the make test was performed, but only in the study by Vanpee et al (21), testers had visual feedback from the monitor to evaluate the quality of the measurement and provided feedback to the patient. Third, length of stay at the ICU before first assessment of muscle strength differed significantly (median 22 d [21] vs median 13 d [22]). Fourth, test positions for knee extension were different between studies: patients were placed supine either with the tested leg in 90° hip flexion and 90° knee flexion or with a bolster under the knee of the tested leg (22). Fifth, Baldwin et al (22) performed ICC calculations based on the average value of three efforts that showed a better reliability compared to calculations with the first, best, or average of the first two efforts. Vanpee et al (21) used the highest score of three efforts (provided that there was less than 10% difference between two measurements). Finally, only Baldwin et al (22) revealed that the time needed to reach the peak force during a maximal voluntary contraction was delayed in the critically ill sample (mean, 4.35 s; SD = 1.05) compared with healthy subjects (mean, 3.75 s; SD = 0.77; $p \leq 0.001$).

The ability to detect changes has also been investigated (22). The SE of measurement and the minimal detectable difference were calculated to determine the minimum amount of change that would be required from a patient's baseline force measurement to detect a real difference in strength. An increase in strength ranging from 14% to 25% of the initial measurement for healthy subjects and from 47% to 110% in critically ill patients would be required to be confident that a true change in muscle strength had occurred beyond measurement error (22).

Furthermore, it has been shown that dynamometry is a more sensitive method to detect changes in strength. Once muscle strength is sufficient to overcome gravity (MRC > 3), changes are better reflected by changes in HHD scores than the MRC (7, 8).

Interrater reliability scores of HGD were rather similar (18, 22). In both studies, the test positions were securely standardized. Furthermore, Hermans et al (18) assessed handgrip strength provided that a minimal MRC score of 3 for both wrist extension and forearm flexion bilaterally was reached. Baldwin et al (22) did not specify this as an inclusion criterion.

Limitations

Volitional muscle strength assessment is associated with several limitations that may challenge the performance in critically ill patients. The cooperation, adequacy, and motivation of the patient are major influencing factors to obtain reliable measurements. In three of six reviewed studies (18, 21, 22), it was reported that muscle strength could only be assessed after a median length of stay in the ICU of approximately 2 weeks due to, among others, the delayed wakening of the patient. To bypass the period where patients are not cooperative, nonvolitional muscle strength assessment, such as electrical and magnetic neuromuscular twitch stimulation, may serve as an alternative (14). Currently, this technique has been successfully applied in critically ill patients for the measurement of the abductor pollicis force (14). However, no studies met the inclusion criteria of this review according to the definition of intra- and interrater reliability. Limitations with this type of testing are expenses, time investment, and requirement of expert knowledge. In addition, normal values are not developed for this technique. Ultrasound provides a measure for limb muscle mass and can also be considered a (nonvolitional) surrogate for muscle strength as was demonstrated in healthy subjects and athletes (25–27). Several studies identified that this measure is sensitive to detect changes in muscle mass in critically ill patients (28, 29). However, the correlation between ultrasound measurements and muscle strength in critically ill patients is unknown. Finally, an alternative assessment that can identify electrophysiological changes and neuromuscular abnormalities can be derived from electromyogram data to diagnose ICUAW (5).

Practical Implications

Assessment of limb muscle strength is important in patients who are at risk in whom ICUAW develops. Previous studies reported a prevalence of ICUAW between 25% and 33% on clinical evaluation using the MRC-sum score (< 48) and up to 58% on electrophysiological evaluation in mechanically ventilated patients for at least 4–7 days (30). Overall, agreement between testers on identification of ICUAW or clinical weakness (MRC-sum score < 48) in critically ill patients is moderate to good. Agreement for identifying severe muscle weakness (MRC-sum score < 36) was excellent (κ 0.93) (18). Lee et al (1) assigned their patients to the quartiles MRC-sum scores (0–15, 16–30, 31–45, and 46–60) and found that participants in the two lowest quartiles had higher mortality rates than those in the two highest quartiles (40% and 16.7% vs 4.2% and 3.5%) (1). Also De Jonghe et al (10, 31) and Ali et al (2) demonstrated that MRC-sum score less than 48 and handgrip strength below 7 kg force for women and 11 kg force for men is associated with prolonged duration of mechanical ventilation, ICU length of stay, hospital length of stay, and in-hospital mortality (2, 10, 31). However, Lee et al (1) did identify the MRC-sum score as an independent predictor for these outcomes but not handgrip strength. The differences may be due to different study populations (medical ICU vs surgical ICU) and the illness severity (Acute Physiology and Chronic Health Evaluation [APACHE] score of 65.8 vs APACHE score of 15.2) (1, 2). HHD has not yet been used to predict outcome.

Only patients who are fully awake and adequate should undergo volitional muscle strength assessment. This means a score of 5 on the five standardized questions or a score of more than 8 of 10 on the Attention Screening Examination scale. In uncooperative patients, nonvolitional measurement of muscle strength with twitch stimulation could be considered. In cooperative patients, ICUAW can be detected with the MRC-sum score (< 48) (10). However, when the goal is to measure progression in muscle strength during recovery of critical illness or in the rehabilitation process, several options for muscle strength assessment are available. In patients who are not strong enough to overcome gravity, the MRC scale is a sensitive method. For patients who are strong enough to overcome gravity (MRC > 3), assessment with dynamometry is more sensitive to detect weakness and progress in muscle strength (7, 32). Due to the variability in HHD in critically ill patients, improvements of approximately 50% are required to reflect a true change in muscle strength (22). Handgrip strength measurements are easily applicable and reliable. However, there is no consistency about the handgrip strength being a valid representative for global muscle strength in critically ill patients (1, 33).

The use of well-defined standardized test positions is critical for the reliability of any muscle strength measurement. This standardization includes positioning of the patient (sitting or supine), the limb position and joint angle for the tested muscle groups, hand position of the tester while applying resistance, contraction time, and verbal encouragement given to the patient. In addition, it also includes the number of repetitions, learning attempts, rest periods in between tests, and information about handling missing data. A protocol for the MRC scale developed by Hermans et al (18) has been added (**supplemental data**, Supplemental Digital Content 1, http://links.lww.com/CCM/A780). A protocol for dynamometry provided by Baldwin et al (22), Vanpee et al (21), and Hermans et al (18) has been added (**supplemental data**, Supplemental Digital Content 2, http://links.lww.com/CCM/A781).

Future research should investigate the ability of manual muscle testing and dynamometry to assist in guiding the rehabilitation process. The sensitivity of manual muscle testing and dynamometry in longitudinal studies and on the relationship of muscle strength with functional outcome needs further attention. Nonvolitional muscle strength measurements in critically ill patients can be useful in early screening of the development of ICUAW. Sensitivity and reliability of nonvolitional muscle strength assessment, such as magnetic or electrical twitch stimulation or ultrasound measurements, require further research.

CONCLUSION

Voluntary muscle strength measurement has proven reliable in critically ill patients provided that strict guidelines on adequacy and standardized test positions are followed.

ACKNOWLEDGMENT

The authors are grateful to Jens de Groot for his advice concerning search strategies.

REFERENCES

1. Lee JJ, Waak K, Grosse-Sundrup M, et al: Global muscle strength but not grip strength predicts mortality and length of stay in a general population in a surgical intensive care unit. *Phys Ther* 2012; 92:1546–1555
2. Ali NA, O'Brien JM Jr, Hoffmann SP, et al; Midwest Critical Care Consortium: Acquired weakness, handgrip strength, and mortality in critically ill patients. *Am J Respir Crit Care Med* 2008; 178:261–268
3. van der Schaaf M, Beelen A, de Vos R: Functional outcome in patients with critical illness polyneuropathy. *Disabil Rehabil* 2004; 26:1189–1197
4. Herridge MS, Tansey CM, Matté A, et al; Canadian Critical Care Trials Group: Functional disability 5 years after acute respiratory distress syndrome. *N Engl J Med* 2011; 364:1293–1304
5. Stevens RD, Marshall SA, Cornblath DR, et al: A framework for diagnosing and classifying intensive care unit-acquired weakness. *Crit Care Med* 2009; 37:S299–S308
6. Bohannon RW: Measuring knee extensor muscle strength. *Am J Phys Med Rehabil* 2001; 80:13–18
7. Bohannon RW: Manual muscle testing: Does it meet the standards of an adequate screening test? *Clin Rehabil* 2005; 19:662–667
8. Schwartz S, Cohen ME, Herbison GJ, et al: Relationship between two measures of upper extremity strength: Manual muscle test compared to hand-held myometry. *Arch Phys Med Rehabil* 1992; 73:1063–1068
9. Kleyweg RP, van der Meché FG, Schmitz PI: Interobserver agreement in the assessment of muscle strength and functional abilities in Guillain-Barré syndrome. *Muscle Nerve* 1991; 14:1103–1109
10. De Jonghe B, Sharshar T, Lefaucheur JP, et al; Groupe de Réflexion et d'Etude des Neuromyopathies en Réanimation: Paresis acquired in the intensive care unit: A prospective multicenter study. *JAMA* 2002; 288:2859–2867
11. Bohannon RW: Reference values for extremity muscle strength obtained by hand-held dynamometry from adults aged 20 to 79 years. *Arch Phys Med Rehabil* 1997; 78:26–32
12. Andrews AW, Thomas MW, Bohannon RW: Normative values for isometric muscle force measurements obtained with hand-held dynamometers. *Phys Ther* 1996; 76:248–259
13. Burns SP, Spanier DE: Break-technique handheld dynamometry: Relation between angular velocity and strength measurements. *Arch Phys Med Rehabil* 2005; 86:1420–1426
14. Mokkink LB, Terwee CB, Patrick DL, et al: The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Qual Life Res* 2010; 19:539–549
15. Harris ML, Luo YM, Watson AC, et al: Adductor pollicis twitch tension assessed by magnetic stimulation of the ulnar nerve. *Am J Respir Crit Care Med* 2000; 162:240–245
16. Mokkink LB, Terwee CB, Gibbons E, et al: Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. *BMC Med Res Methodol* 2010; 10:82
17. Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977; 33:159–174
18. Hermans G, Clerckx B, Vanhullebusch T, et al: Interobserver agreement of Medical Research Council sum-score and handgrip strength in the intensive care unit. *Muscle Nerve* 2012; 45:18–25
19. Hough CL, Lieu BK, Caldwell ES: Manual muscle strength testing of critically ill patients: Feasibility and interobserver agreement. *Crit Care* 2011; 15:R43
20. Fan E, Ciesla ND, Truong AD, et al: Inter-rater reliability of manual muscle strength testing in ICU survivors and simulated patients. *Intensive Care Med* 2010; 36:1038–1043
21. Vanpee G, Segers J, Van Mechelen H, et al: The interobserver agreement of handheld dynamometry for muscle strength assessment in critically ill patients. *Crit Care Med* 2011; 39:1929–1934
22. Baldwin CE, Paratz JD, Bersten AD: Muscle strength assessment in critically ill patients with handheld dynamometry: An investigation of reliability, minimal detectable change, and time to peak force generation. *J Crit Care* 2013; 28:77–86
23. de Vet HC, Adèr HJ, Terwee CB, et al: Are factor analytical techniques used appropriately in the validation of health status questionnaires? A systematic review on the quality of factor analysis of the SF-36. *Qual Life Res* 2005; 14:1203–1218; discussion 1219
24. Shrout PE, Fleiss JL: Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 1979; 86:420–428
25. Reed RL, Pearlmutter L, Yochum K, et al: The relationship between muscle mass and muscle strength in the elderly. *J Am Geriatr Soc* 1991; 39:555–561
26. Maughan RJ, Watson JS, Weir J: Strength and cross-sectional area of human skeletal muscle. *J Physiol* 1983; 338:37–49
27. Payette H, Hanusaik N, Boutier V, et al: Muscle strength and functional mobility in relation to lean body mass in free-living frail elderly women. *Eur J Clin Nutr* 1998; 52:45–53
28. Reid CL, Campbell IT, Little RA: Muscle wasting and energy balance in critical illness. *Clin Nutr* 2004; 23:273–280
29. Gruther W, Benesch T, Zorn C, et al: Muscle wasting in intensive care patients: Ultrasound observation of the ***M. quadriceps*** femoris muscle layer. *J Rehabil Med* 2008; 40:185–189
30. Hermans G, De Jonghe B, Bruyninckx F, et al: Clinical review: Critical illness polyneuropathy and myopathy. *Crit Care* 2008; 12:238
31. De Jonghe B, Bastuji-Garin S, Sharshar T, et al: Does ICU-acquired paresis lengthen weaning from mechanical ventilation? *Intensive Care Med* 2004; 30:1117–1121
32. Bohannon RW: Is it legitimate to characterize muscle strength using a limited number of measures? *J Strength Cond Res* 2008; 22:166–173
33. Bohannon RW: Hand-grip dynamometry predicts future outcomes in aging adults. *J Geriatr Phys Ther* 2008; 31:3–10