A Standardized Approach to the Fugl-Meyer Assessment and Its Implications for Clinical Trials

Jill See, MPT\textsuperscript{1}, Lucy Dodakian, MA\textsuperscript{1}, Cathy Chou, MPT\textsuperscript{1}, Vicky Chan, MSPT\textsuperscript{1}, Alison McKenzie, PhD\textsuperscript{2}, David J. Reinkensmeyer, PhD\textsuperscript{1} and Steven C. Cramer, MD\textsuperscript{1}

Abstract

Background. Standardizing scoring reduces variability and increases accuracy. A detailed scoring and training method for the Fugl-Meyer motor assessment (FMA) is described and assessed, and implications for clinical trials considered. Methods. A standardized FMA scoring approach and training materials were assembled, including a manual, scoring sheets, and instructional video plus patient videos. Performance of this approach was evaluated for the upper extremity portion. Results. Inter- and intrarater reliability in 31 patients were excellent (intraclass correlation coefficient = 0.98-0.99), validity was excellent ($r = 0.74-0.93$, $P < .0001$), and minimal detectable change was low (3.2 points). Training required 1.5 hours and significantly reduced error and variance among 50 students, with arm FMA scores deviating from the answer key by $3.8 \pm 6.2$ points pretraining versus $0.9 \pm 4.9$ points posttraining. The current approach was implemented without incident into training for a phase II trial. Among 66 patients treated with robotic therapy, change in FMA was smaller ($P \leq .01$) at the high and low ends of baseline FMA scores. Conclusions. Training with the current method improved accuracy, and reduced variance, of FMA scoring; the 20% FMA variance reduction with training would decrease sample size requirements from 137 to 88 in a theoretical trial aiming to detect a 7-point FMA difference. Minimal detectable change was much smaller than FMA minimal clinically important difference. The variation in FMA gains in relation to baseline FMA suggests that future trials consider a sliding outcome approach when FMA is an outcome measure. The current training approach may be useful for assessing motor outcomes in restorative stroke trials.

Keywords

stroke, motor, outcome measures, validity, reliability, training

Introduction

Motor deficits are among the most common after stroke\textsuperscript{1,2} and thus are a major contributor to stroke disability. A number of restorative therapies have been examined for the ability to improve motor outcome.\textsuperscript{3} A key factor in assessing therapeutic efficacy is the choice of behavioral outcome measure. In restorative stroke studies, as with acute stroke, outcome measures need to be valid, reliable, and responsive to change.\textsuperscript{4,5}

A number of clinical trials focused on motor outcomes after stroke have employed the Fugl-Meyer Motor Assessment (FMA). The FMA was designed by Fugl-Meyer et al\textsuperscript{6} to provide a numeric score of motor status after stroke based on the sequential stages of motor recovery described by Twitchell,\textsuperscript{7} Reynolds et al,\textsuperscript{8} and Brunstrom\textsuperscript{9} using measures such as limb synergy and range of motion.\textsuperscript{6} The FMA has been found to be valid\textsuperscript{10,11} and reliable.\textsuperscript{6,11-13} The FMA has received increasing attention as a clinical trial outcome measure, in part because it has been found to be sensitive to behavioral gains in the setting of wide-ranging interventions that include pharmacological compounds,\textsuperscript{14,15} robotics,\textsuperscript{16,17} brain stimulation,\textsuperscript{18,19} constraint-induced therapy,\textsuperscript{20} neuromuscular stimulation,\textsuperscript{21} mental practice,\textsuperscript{22} and virtual reality therapy.\textsuperscript{23} Consistent with this, a search of PubMed using “Fugl-Meyer” and “clinical trial” yields 200 references; searching with the same 2 terms on www.clinicaltrials.gov yields 125 studies.

A number of issues complicate use of the FMA in clinical trials, however, and motivate the current report. First, the FMA was described as having the goal to standardize
patient assessment. However, key operational details of test administration were not included by Fugl-Meyer et al, raising uncertainty as to consistency of FMA scores across sites and time. Two publications have outlined protocols for scoring FMA24,25; however, omission of key operational details for administering the test and for scoring patient responses limits the extent to which these protocols standardize patient assessment with the FMA. The importance of standardizing methods for scoring outcome measures in stroke trials continues to be emphasized26-28; training materials are needed toward this goal, a perspective that was the underlying goal of the current report. Second, the structure of the FMA creates uncertainties regarding its use in a clinical trial setting. For example, the FMA is not truly linear—a gain of points at the bottom versus the top end of the scale has different meaning. Also, prior studies have been divided as to the extent of floor and ceiling effects with the FMA.11,29-31 As a result, it is unclear whether the FMA performs equally well as an outcome measure across the full spectrum of poststroke motor deficits, and how the FMA performs as a stratiﬁering variable.32

These issues are addressed in the current report. First, detailed definitions and training methods for scoring the FMA in a standardized manner are presented, based directly on the method described by Fugl-Meyer et al3; note that the instructional video, test patient videos, and standardized scoring sheets associated with this approach are provided in the supplementary material available at http://nnr.sagepub.com/content/by/supplemental-data. The current approach was created and evaluated during 2006-2009, prompted by observation of differences in FMA scores across examiners on our team. Second, for the upper extremity, the reliability, minimal detectable change (MDC), and validity of this testing approach are described herein. Performance of the total FMA, proximal arm subsection, and distal arm (wrist/hand) subsection are each examined separately, to increase the granularity of the FMA for detecting differences between patients. Third, the effect of training with this approach is examined, including effect on precision and on variance of FMA scoring, with consideration of power and sample size in a clinical trial context. Fourth, the performance of FMA score as a stratiﬁering variable was assessed in patients receiving arm motor robotic therapy. Fifth, the experience of implementing this scoring approach in a phase II clinical trial of stroke recovery is reviewed.

Methods

Design and Content of the Current Approach

The FMA consists of 33 items for the upper extremity and 17 for the lower extremity. Each item is scored on a 3-point ordinal scale (0, 1, or 2), with 0 generally corresponding to no function, 1 to partial function, and 2 to perfect function. The items are summed to provide a final score, with maximum score (no impairment) of 66 points for the upper extremity and 34 points for the lower extremity.

In the FMA testing approach presented herein, one side of the body is tested/scored, and then the other body side is tested/scored. For each test item, the initial subject limb position is described, testing materials are listed, specific instructions to be read to the patient are provided, specific assessor movements and amount of assistance that may be provided are outlined, then the specific details by which each item is scored are provided. The score is based on best performance. The task is to be performed within a reasonable time frame, with 20 seconds per attempt used as a cut-off based on experience, and a maximum of 3 attempts per test item. No special considerations in scoring are made for presence of amputation, contracture, prosthesis, aphasia, or orthopedic problems.

A total of 10 items were generated in support of this FMA scoring approach, each of which is provided in the online supplementary material available at http://nnr.sagepub.com/content/by/supplemental-data:

1. Arm FM reference manual: Provides the details of the approach for the arm, useful to have at one’s side when formally assessing patients on the FMA
2. Arm FM scoring sheet: Provides a standardized sheet for scoring the arm FMA; printing multiple copies for formally assessing patients on the FMA is suggested
3. Arm FM training video: Provides video examples of a patient being scored on the arm FMA, part of training to standardize FMA scoring
4. Arm FM training video guide: Provides written explanations to assist with the video
5. Arm FM test subject 1 video: A patient with stroke (arm FM score = 30) is to be assessed by trainees while watching this video, can be used as part of FMA training
6. Arm FM test subject 1 answer key: Reviewed after assessing the subject 1 video, this key provides correct scores and explanations of scoring as determined by the 5 experienced therapists (JS, LD, CC, VC, AMcK) of this project
7. Arm FM test subject 2 video: A second patient with stroke (arm FM score = 46) is to be assessed by trainees while watching, also useful as part of FMA training
8. Arm FM test subject 2 answer key: Reviewed after assessing the subject 2 video, this key provides correct scores and explanations of scoring as determined by the 5 experienced therapists (JS, LD, CC, VC, AMcK) of this project
9. **Leg FM reference manual**: Provides the details of the approach for the leg
10. **Leg FM scoring sheet**: Provides a standardized sheet for scoring the leg FMA

The 2 patients were selected for the videos because they have mid-range scores and because each permits demonstration of several key distinctions in FMA scoring. Answer key scores for these two patients were based on the FMA testing approach described herein, with maximum adherence to the principles described in the original FMA article, and were derived from group discussions among study therapists. Two training components related to the leg (items 9 and 10 above) are included, although current analyses are focused on components related to the arm (items 1-8).

The recommended standardized arm FMA training procedure proposed herein is to (a) review the reference manual and scoring sheet, (b) watch the subject 1 video and score this patient, (c) watch the training video with the video guide at hand, (d) watch the subject 2 video and score this patient, and then (e) review the two answer keys and note scoring discrepancies. The time to complete this training procedure is estimated to be 1.5 hours.

All studies described in the current report were approved by the University of California, Irvine, Institutional Review Board. All patients provided informed consent.

### Reliability, Minimal Detectable Change, and Validity

Patients were recruited from 2 robotic studies of arm motor therapy that enrolled adults with a chronic (>3 months prior) stroke, arm motor deficits, and no severe deficits in language or attention. One of the 2 studies required that the motor deficits be right-sided. A total of 31 patients were recruited, providing 80% power at α = .05 to detect an intraclass correlation coefficient (ICC) of 0.91. The reliability and validity evaluations were shared across 2 physical therapists and an occupational therapist, each of whom had at least 10 years experience in assessing stroke patients and each of whom also had participated in creating the current FMA manual and training materials. Reliability assessments were obtained prior to initiating robotic therapy. For intrarater reliability, each of the 31 patients received 2 exams from one of the therapists, 1 week apart; for 4 patients, a second therapist was available to provide an additional independent intrarater reliability assessment, bringing the total number of intrarater assessments to a total of 35. For interrater reliability, patients underwent 2 independent exams, performed by different therapists, separated by <1 hour. Four patients were unable to undergo this added testing, bringing the total number of intrarater assessments to 27.

To assess FMA validity, at one of the sessions, an examiner also scored 6 other motor-related assessments: grip strength, pinch strength, Box & Blocks, the Action Research Arm Test (ARAT), 9-hole peg, and the Stroke Impact Scale (SIS) hand subscore (SIS II, Q 7a-7e). These 6 were selected to capture the diverse dimension of stroke effects on the arm motor system, and include tests of body function and of activities limitations, tests that are patient-reported as well as tests that are examiner-based, and tests of the proximal arm and of the distal arm. These exams were performed on 12 of the patients, at each of 4 separate visits across the treatment period, for a total of 48 exams focused on validity.

### Effect of Training on Accuracy and Variance of Fugl-Meyer Assessment Scoring

Students (n = 50) in the Chapman University DPT program were trained in the arm FMA using the recommended arm FMA training approach described above. Thus, each student scored a video of a stroke patient before and again after undergoing training on the arm FMA using the current approach. Pretraining, each student watched one of the Arm FM test subject videos (half watching subject 1 first, half watching subject 2 first) and completed an arm FM scoring sheet, pausing the video after each task for additional time to score, if needed. Next, each student was trained by watching the Arm FM training video with the Arm FM training video guide at hand, and reviewed the Arm FM reference manual. Finally, each student was then tested posttraining by watching the remaining Arm FM test subject video (eg, watched subject 2 video posttraining if the subject 1 video had been completed pretraining) then completing a second Arm FM scoring sheet, pausing as needed. At the end of scoring, students were provided with items 6 and 8 above (ie, the Arm FM test subject 1 answer key and the Arm FM test subject 2 answer key), and in this way the students’ scores were compared with the correct scores that were determined by the 5 experienced therapists (JS, LD, CC, VC, AMcK) of this project.

### Ability of the Fugl-Meyer Assessment to Perform as a Stratifying Variable

In 66 patients with chronic stroke enrolled in 1 of 3 studies using robotic therapy to improve arm motor function in chronic stroke (clinicaltrials.gov identifier: NCT01244243), arm motor FMA scores were measured at baseline and 1 month after a course of robot therapy. These studies enrolled adults with a chronic (>3 months prior) stroke, arm motor deficits, and no severe deficits in language or attention. FMA scores were obtained by 1 of the 5 experienced therapists (JS, LD, CC, VC, AMcK) of this project. The ability of
baseline FMA score to stratify patients according to treatment gains was examined by studying the relationship between FMA baseline and change scores.

**Implementation in a Clinical Stroke Trial**

All 10 training materials were provided to the 11 United States, Canada, and Germany enrollment sites as part of pretrial training for the clinical stroke trial “A Single-Blind Study of the Safety, Pharmacokinetics and Pharmacodynamics of Escalating Repeat Doses of GSK249320 in Patients With Stroke” (clinicaltrials.gov identifier NCT00833989), which compared GSK249320, a humanized monoclonal antibody that neutralizes MAG (myelin-associated glycoprotein)-mediated inhibition, with placebo. The FMA was among the secondary endpoints of this safety trial. Sites were asked to review the Arm FM reference manual, Arm FM scoring sheet, Arm FM training video, and Arm FM training video guide as part of trial participation. This part of the study examined the feasibility of broad implementation of the current FMA training procedure. The measure examined in relation to this goal was the number of problems or issues that were reported to the study sponsor in relation to FMA training.

**Statistics**

Two-tailed statistical methods were used and, except for the polynomial statistics in the stratification analyses, were nonparametric. Intrar- and interrater reliabilities were assessed using Spearman’s rank order correlation, ICC, and MDC. MDC was assessed as the MDC_{90}, which indicates the 90% confidence interval that the magnitude of measurement variability will be less than the MDC values. Validity was measured by comparing the FMA score with each of the 6 motor-related assessments of interest (grip strength, pinch strength, Box & Blocks, ARAT, 9-hole peg, and SIS hand subscore), using Spearman’s rank order correlation. Reliability and validity were each evaluated for the total arm FMA, proximal arm portion of the FMA, and distal arm (hand/wrist) portion of the FMA. For analysis of the standardization procedures, differences between students’ scores and the answer key pretraining were evaluated for significance by performing a 1-sample Wilcoxon signed-rank test to determine whether these differences were significantly different from zero. This was then repeated for values obtained posttraining. The stratification study examined the relationship that total arm FMA score at baseline had with the change in total arm FMA score across the period of therapy. Initial analysis tested for a linear relationship using Spearman’s rank order correlation. Visual inspection of the data suggested a second-order relationship, which was tested using a quadratic (polynomial) fit. Post hoc testing extended this observation by testing whether change in total arm FMA score across therapy differed significantly according to baseline score using the Wilcoxon rank sums test. Analyses used JMP 5.0 (SAS, Cary, NC). ICC analyses used online tools (http://department.obg.cuhk.edu.hk/researchsupport/IntraClass_correlation.asp).

**Results**

**Reliability and Validity**

Characteristics for the 31 patients appear in Table 1.

For the total arm FMA, reliability (Table 2) was excellent using the scoring approach described herein. For intrarater reliability, Spearman’s $r = 0.99$ ($P < .0001$), ICC = 0.99, and MDC_{90} = 3.2 points (4.8% of the highest possible score). For interrater reliability, $r = 0.97$ ($P < .0001$),

<table>
<thead>
<tr>
<th>Table 1. Stroke Patients Assessed in Reliability Studies.</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
</tr>
<tr>
<td>Age in years (mean ± SD)</td>
</tr>
<tr>
<td>Gender (%)</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Affected side (%)</td>
</tr>
<tr>
<td>Right</td>
</tr>
<tr>
<td>Left</td>
</tr>
<tr>
<td>Handedness</td>
</tr>
<tr>
<td>Right</td>
</tr>
<tr>
<td>Left</td>
</tr>
<tr>
<td>Time poststroke in months (mean ± SD)</td>
</tr>
<tr>
<td>Total NIHSS score; median (IQR)</td>
</tr>
<tr>
<td>Baseline FMA score (mean ± SD)</td>
</tr>
</tbody>
</table>

Abbreviations: SD, standard deviation; NIHSS, National Institutes of Health Stroke Scale; IQR, interquartile range; FMA, Fugl-Meyer Assessment.
Table 2. Reliability Studies.

<table>
<thead>
<tr>
<th>Test</th>
<th>FMA Total Score</th>
<th>FMA Proximal Subscore</th>
<th>FMA Hand/Wrist Subscore</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intrarater reliability</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spearman’s $r^a$</td>
<td>0.99</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>ICC</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>MDC$_{90}$</td>
<td>3.2 points</td>
<td>1.7 points</td>
<td>1.7 points</td>
</tr>
<tr>
<td><strong>Interrater reliability</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spearman’s $r^a$</td>
<td>0.97</td>
<td>0.95</td>
<td>0.85</td>
</tr>
<tr>
<td>ICC</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>MDC$_{90}$</td>
<td>3.2 points</td>
<td>1.6 points</td>
<td>2.5 points</td>
</tr>
</tbody>
</table>

Abbreviations: FMA, Fugl-Meyer Assessment; ICC, intraclass correlation coefficient; MDC$_{90}$, 90% confidence interval that the magnitude of measurement variability is less than the minimal detectable change.

$^a$All $r$ values are Spearman’s rank order correlation, and in all cases were significant with $P < .0001$.

Table 3. Validity Studies.$^a$

<table>
<thead>
<tr>
<th>Test</th>
<th>Baseline Value</th>
<th>Correlation With FMA Total Score</th>
<th>Correlation With FMA Proximal Subscore</th>
<th>Correlation With FMA Hand/Wrist Subscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum grip force, affected/nonaffected hand</td>
<td>0.29 ± 0.22</td>
<td>0.74</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>Maximum pinch force, affected/non-affected hand</td>
<td>0.40 ± 0.30</td>
<td>0.88</td>
<td>0.87</td>
<td>0.85</td>
</tr>
<tr>
<td>Box &amp; Blocks (no. of blocks)</td>
<td>0 (0-29)</td>
<td>0.86</td>
<td>0.79</td>
<td>0.88</td>
</tr>
<tr>
<td>ARAT score</td>
<td>27.2 ± 22.5</td>
<td>0.93</td>
<td>0.88</td>
<td>0.89</td>
</tr>
<tr>
<td>9-hole peg (no. of pegs placed)</td>
<td>0 (0-7)</td>
<td>0.75</td>
<td>0.64</td>
<td>0.80</td>
</tr>
<tr>
<td>SIS hand subscore</td>
<td>2.3 ± 1.3</td>
<td>0.86</td>
<td>0.79</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Abbreviations: FMA, Fugl-Meyer Assessment; ARAT, Action Research Arm Test; SIS, Stroke Impact Scale.

$^a$Baseline values are mean ± SD except for Box & Blocks and 9-hole peg scores, which are median (interquartile range). Spearman $r$ values are presented for correlation values, which in all cases were significant with $P < .0001$. Mean FMA score across exams was 38 ± 15 points.

ICC = 0.99, and MDC$_{90}$ = 3.2 points (which is 4.8% of the highest possible score).

For the portion of the FMA corresponding to proximal arm motor function, for intrarater reliability, $r = 0.99$ ($P < .0001$), ICC = 0.99, and MDC$_{90}$ = 1.7 points (4.7% of the highest possible score). For interrater reliability, $r = 0.95$ ($P < .0001$), ICC = 0.98, and MDC$_{90}$ = 1.6 points (4.4% of the highest possible score).

For the portion of the FMA corresponding to hand/wrist (distal arm) motor function, for intrarater reliability, $r = 0.94$ ($P < .0001$), ICC = 0.99, and MDC$_{90}$ = 1.7 points (7.1% of the highest possible score). For interrater reliability, $r = 0.85$ ($P < .0001$), ICC = 0.98, and MDC$_{90}$ = 2.5 points (10.4% of the highest possible score).

Validity was also excellent (Table 3), for example, Spearman’s $r$ values for the total arm FMA score ranged from 0.74 to 0.93. The total arm FMA score, FMA proximal subscore, and FMA hand/wrist subscore each correlated significantly ($P < .0001$ in all cases) with each of the diverse motor assessments, including patient-reported outcomes (SIS hand subscore), distal motor function (9-hole peg, grip force, and pinch), and combined distal and proximal (Box & Blocks and ARAT) assessments.

**Effect of the Standardized Training Procedure on Accuracy and Variance of Fugl-Meyer Scoring**

Of the 50 students, a posttesting survey disclosed that only one had ever administered the arm FMA previously, on a single occasion. Participation in the standardized training procedure significantly improved accuracy. Pretraining, the difference between students and the answer key score was 3.8 ± 6.2 (mean ± SD) points for the total arm FMA (students underscored compared with the answer key), 2.6 ± 3.8 for the proximal FMA (students underscored), and 0.1 ± 3.8 for the wrist/hand FMA (students overscored). The total arm FMA and proximal FMA values were each significantly different from zero ($P < .0001$), indicating that the students deviated significantly from the answer key. Posttraining, the difference between students and the answer key score was 0.9 ± 4.9 points for the total arm FMA (a 20% decrease in SD), 0.7 ± 2.8 for the proximal FMA, and 0.16
± 3.2 for the wrist/hand FMA; none of these was significantly different from zero, indicating that the students showed no significant differences from the answer key after training.

**Ability of the Fugl-Meyer Assessment to Perform as a Stratifying Variable**

The 66 patients had age 59.3 ± 14.1 years, time poststroke 17 ± 34.6 months, gender 43 males/23 females, baseline total arm FMA 36.9 ± 14.5 points, and change in FMA 3.8 ± 3.6 points. Baseline total arm FMA score was not linearly related to change in FMA score across arm robotic therapy ($r = 0.15, P > .2$); however, the second-degree polynomial relationship between these 2 measures was significant ($r = 0.47, P = .0001$, Figure 1A), indicating a U-shaped relationship. Consistent with this, the change in total arm FMA across therapy was significantly smaller for subjects with the lowest (FMA < 20) and for subjects the highest (FMA > 55) baseline scores, as compared with the middle ($P = .01$ to $P = .0002$, see Figure 1B).

**Implementation in a Clinical Stroke Trial**

All 11 sites received the training materials and completed the requested FMA training prior to subject enrollment. There were no problems or issues reported at any study site with the implementation of this standardized FMA training procedure.

**Discussion**

A number of therapies are under study to improve motor outcome after stroke. The FMA has been a common choice for assessing treatment effects in this setting. A need exists for a detailed approach to FMA scoring to maximize consistency over time and across sites, and for training materials for such a scoring approach. The current report describes such a detailed approach to FMA scoring (presented in full in the supplementary material available at http://nnr.sagepub.com/content/by/supplemental-data), then reviews experience with FMA training materials based on this approach. This approach was found to be reliable (Table 2) and valid.
improve patient compliance; furthermore, one recent ability to monitor care, facilitate communication, and reported outcomes, which have received attention for their when FMA validity was measured in relation to patient-values 0.98 to 0.99, similar to those of Sullivan et al24 and current method was found to be reliable (Table 2), with ICC method and training procedure address these concerns. The which stroke affects the motor system (Table 3), was also motor assessments that considered the many dimensions by timeless.39 Prior protocols for measuring FMA have been pub-
lished,24,25 but these did not always specify details critical to standardized FMA scoring, such as the exact instructions that are provided to patients, initial limb position for testing, or the amount of assistance that the examiner can provide; did not include full training materials such as videos (see supplementary material available at http://nnr.sagepub.com/content/by/supplemental-data); and have been criti-
cized for introducing modifications of the original Fugl-Meyer scale.44 The current standardized FMA scoring method and training procedure address these concerns. The current method was found to be reliable (Table 2), with ICC values 0.98 to 0.99, similar to those of Sullivan et al24 and Platz et al.45 Validity, examined in relation to wide-ranging motor assessments that considered the many dimensions by which stroke affects the motor system (Table 3), was also excellent and similar to prior findings.45 This remained true when FMA validity was measured in relation to patient-reported outcomes, which have received attention for their ability to monitor care, facilitate communication, and improve patient compliance; furthermore, one recent study found that the SIS, the patient-reported outcome measure examined in the current study (Table 3), can be a unique source of insight into arm motor status after stroke.39 Reliability and validity findings were true for the total arm FMA, the FMA proximal subscore, and the FMA wrist/hand (distal) subscore. The proximal FMA and the distal FMA subscores were included in analysis because some treatments selectively target the proximal or distal upper extremity. Thus, whereas proximal and distal arm FMA measures may not show separate dimensions in the setting of spontaneous stroke recovery, assessment of proximal arm separately from distal arm may be important to best capture or predict effects of some treatments. The training procedure improved consistency and accuracy of scoring. Together these findings support the utility of the current approach for measuring FMA.

The current findings have direct implications for clinical trials given that the current FMA standardized training procedure significantly reduced variance and improved accuracy in FMA scoring. Specifically, training was associated with a 20% decrease in the SD for the total arm FMA score. This level of reduction in variance of outcome measure scoring is not trivial: for a trial powered at 80% to detect a difference of 7 points on the total arm FMA between 2 treatment groups, with \( \alpha = .05 \), and with baseline SD of 14.5 (as in Table 1), a 20% reduction in variance would cut the total sample size needed to enroll from 137 to 88. The current FMA standardized training procedure improved accuracy: the students’ total arm FMA scores before training were significantly different from answer key values, whereas students’ scores after training did not differ from the answer key. The significant departure of FMA scores from the correct value prior to training would obfuscate detection of treatment effects. The MDC data provide further support that training with this approach is important to clinical trials. The MDC\(_{90}\) for total arm FMA score using the current method was 3.2 points, indicating that a change of total arm FMA score greater than 3.2 points for an individual is necessary to be 90% certain that the change is not because of measurement error. This value is substantially smaller than estimates for minimal clinically important difference for total arm FMA, which include 7.25 points (ie, half of a SD, across scales and populations), 6.6 points (ie, 10% of the range for any given scale), 10.6 to 19.8 points (ie, 16% to 30% of the range, determined with respect to various upper extremity assessments in patients receiving inpatient rehabilitation 10-26 days after stroke), 4.25 to 7.25 points (in patients an average of 59 months poststroke enrolled in a clinical trial, and depending on which aspect of motor function is used for comparison), and 10 points (in patients receiving inpatient rehabilitation 17 days after stroke). Clearly, for the FMA, as with many other measures, minimal clinically important difference is context dependent, for example, varying with the method used to define clinical significance or with the population under study (eg, what constitutes a clinically important change is different 1 week vs 1 year after stroke). The 3.2 point MDC\(_{90}\) value is also smaller than many treatment effects measured by the total arm FMA, for example, 25 points with amphetamine initiated 8 days after stroke,14 34 points with fluoxetine initiated 9 days after stroke,15 or 8 to 9 points in studies of robotic therapy enrolling patients months/years after stroke onset.16,17

The change in total arm FMA score associated with robot therapy had a second-order (quadratic), rather than a linear,
relationship with baseline FMA score (Figure 1), a finding that informs the use of the FMA as an endpoint in studies enrolling patients with a broad range of motor deficits. There are several possible explanations for this finding, including that the FMA is not truly linear. Fugl-Meyer et al2 based FMA scoring on the sequence of stages of spontaneous recovery,6-9 but this succession is not truly linear, and so a FMA increase from 10 to 20 points does not necessarily have the same meaning as an increase from 50 to 60 points. Also, the FMA may have a floor and ceiling effect. The literature is divided on this point,10,29-31 possibly because it might be more true in some contexts more than others. Figure 1 suggests floor and ceiling effects may be present when using the FMA to measure change associated with robotic therapy. Regardless of the explanation, the lack of a linear relationship between baseline total arm FMA score and the FMA score change with treatment suggests that MDC and minimal clinically important difference might vary according to the population under study, a perspective arising frequently in studies of stroke given the heterogeneity of this population.

How might future trials using FMA as an endpoint build on the current finding that the relationship between baseline FMA and change in FMA is quadratic not linear? For trials that evaluate treatment response dichotomously, that is, as successful or not, one potential response is to define a successful outcome in a manner that varies with baseline status. This approach is known as a sliding dichotomous outcome, or responder analysis. A recent analysis of acute stroke trial outcome measures emphasized the utility of this approach,59 and noted its ability to increase study power. With this approach, patient subgroups are specified before the trial on the basis of established prognostic measures such as age, baseline behavioral status, or extent of injury. Successful response to therapy is defined differently for each subgroup. A sliding outcomes approach has been used in several acute stroke trials.59 For example, the ABESTT-II trial of Abciximab for acute stroke defined good outcome as modified Rankin Scale (mRS) score of 0 for patients with baseline National Institutes of Health Stroke Scale (NIHSS) score of 4 to 7, mRS score of 0 to 1 for baseline NIHSS of 8 to 14, and mRS score of 0 to 2 for baseline NIHSS 15 to 22.60 A sliding outcomes approach has also been used in the chronic stroke. For example, the LEAPS trial of locomotor training61 defined success in the primary outcome measure (proportion of participants with improved functional walking level) as gait velocity ≥0.4 m/s for enrollees with baseline gait velocity <0.4 m/s and as gait velocity ≥0.8 m/s for enrollees with baseline gait velocity 0.4 to 0.8 m/s.

The value and logic of using a sliding dichotomous (responder) analysis in the context of arm motor recovery is readily appreciated—return of rapid dexterous hand movements might be extremely unlikely in a patient with severe arm motor deficits at baseline, but a boost in grip force of a mere 20 N might be attainable and indeed relevant to function, whereas the same 20 N boost in grip force might be near trivial for a patient with mild baseline deficits. The current results (Figure 1) suggest utility for sliding dichotomous outcomes in clinical trials targeting arm motor function. In such trials, patients at the extremes of arm motor function might define therapeutic success using a different FMA cutoff—or perhaps even using a different scale—as compared with patients with intermediate levels of arm motor function. The exact choice of cutoffs may vary depending on the population and intervention under study. An alternative approach for dealing with the heterogeneity in stroke populations is to use a composite endpoint, as was employed in the Everest trial of epidural motor cortex stimulation,62,63 where the primary outcome measure combined the impairment-based FMA with a second scale (Arm Motor Ability Test) that measured function. The FMA has limited sensitivity to motor-related measures such as executive control, timing, and imagery, and so the choice of a second scale for a composite endpoint might be guided by the content of the therapeutic intervention.

The current report described then assessed an approach to maximize utility of the FMA for stroke recovery. Limitations of the study include that validity and reliability of the current FMA approach might vary across different populations of patients, such as those with severe aphasia or severe neglect or very mild motor deficits, or across different examiners. The reliability data, as in any study of neurologically infirm populations, must be interpreted in light of the potential influence of factors such as fatigue, medication effects, and confusion. Also, the current method standardized training procedure has not been validated outside the English language. The current approach increases accuracy of FMA scoring and so would be expected to improve precision and statistical power in clinical trials that use the FMA as an endpoint.

Acknowledgment
We thank Lisbeth Jääskö for her assistance and guidance.

Declaration of Conflicting Interests
The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Dr Cramer has received grant and consulting fees from GlaxoSmithKline, and consulting fees from Pfizer/Cogstate and MicroTransponder.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by funds provided by the National Center of Research Resources, 5M01 RR-00827-29 and NS059909, US Public Health Service.
References


