Clinimetric properties of the Fugl-Meyer assessment with adapted guidelines for the assessment of arm function in hemiparetic patients after stroke

Satoru Amano1, Atsushi Umeji5, Akira Uchita6, Yukihisa Hashimoto6, Takashi Takebayashi1b, Kayoko Takahashi6c, Yuki Uchiyama6ab and Kazuhiro Domen6b

1Department of Physical Medicine and Rehabilitation, Hyogo College of Medicine Hospital, Nishinomiya, Japan; 2Department of Rehabilitation Medicine, Hyogo College of Medicine, Nishinomiya, Japan; 3Department of Occupational Therapy, Kitasato University, Kanagawa, Japan

ABSTRACT

Background: Against the background of linguistic and cultural differences, there is a need for translation and adaptation from the English version of the Fugl-Meyer Assessment (FMA) to Japanese. In addition, there is no study of inter-rater reliability of FMA all domains for affected upper extremities with appropriate sample size based on the intraclass correlation coefficient (ICC) focusing on non-simultaneous assessment.

Objective: This study aimed (1) to translate the English version of the FMA and its administration/scoring manual; and (2) to investigate the psychometric properties of the Japanese version of the FMA in patients with stroke.

Methods: A prospective single-center study involving 30 patients was conducted. The FMA and the Action Research Arm Test, the Box-and-Block Test, and the Motor Activity Log were employed. The inter-rater/intra-rater reliability, the internal consistency, the validity, and the floor/ceiling effects were assessed.

Results: Regarding the non-simultaneous and simultaneous inter-rater reliability, ICC ranged from 0.809–0.983 (P<0.001) and 0.991–0.999 (P<0.001), respectively. Regarding the simultaneous intra-rater reliability, ICC ranged from 0.994–0.999 (P<0.001). The Cronbach’s alpha was 0.973 in the non-simultaneous evaluation and 0.981 in the simultaneous evaluation. Regarding the validity, Spearman’s rhos were higher than 0.92 for the FMA all domains and motor domain. The patients who showed the highest score and the lowest score of the FMA (all domains and motor domain) were 10% and 0%, respectively.

Conclusions: The Japanese version of the FMA motor domain and all domains can reliably assess the affected upper extremities in patients with mild-to-severe hemiparesis after stroke for both non-simultaneous and simultaneous assessment.

Introduction

As stated in the Guidelines for Adult Stroke Rehabilitation and Recovery (American Heart Association/American Stroke Association Guidelines), formal standardized and validated measures should be used to the extent possible in the rehabilitative care of adults recovering from stroke.1 Regarding the importance of evaluating functional outcome measures, van der Putten et al. also noted, “Measuring the effectiveness of clinical interventions by using standardized measurement instruments is now widely accepted as being central to good clinical practice”.2 A three-stage selection strategy developed by Baker et al. revealed the importance of the Fugl-Meyer Assessment (FMA) for the affected upper extremity after stroke based on clinical content and study design issues.3 The FMA motor tasks are based on the idea of consecutive steps of recovery from hemiparesis after stroke and ordered according to presumed stage of recovery.4,5 The advantages of the FMA is that previous studies have not only showed good reliability, validity, and responsiveness,6 but also high feasibility.7 A recent study of upper extremity outcome measures showed that the FMA was the most frequently used outcome measure in stroke rehabilitation.7

In English-speaking countries, the FMA has shown good reliability,8–11 responsiveness,8,12 and adequate concurrent validity11,13 when compared with other upper extremity measures. In non-English-speaking countries, several researchers have conducted translation and evaluation studies aimed at using the FMA in their own languages.14,15 The psychometric properties of a measurement depend on the population and setting in which it is used, and further tests should be conducted on the psychometric properties of the adapted version after the translation is complete.16 Lundquist et al.15 actually followed the concept of the guidelines for the process of cross-cultural adaptation,16 and the Danish version of the FMA was provided publicly.

Japan is one of those countries in which Western culture is not dominant, and Japanese is the language which did not develop from Indo-European.17 Against the background of linguistic and cultural differences, there is a need for translation and adaptation from the English version of the FMA to Japanese. In addition to being standardized, whether the clinical properties of the Japanese version of the FMA are appropriate to the level of clinical trial use should be confirmed. Therefore, the present study has two purposes: (1) translation of the English version...
of the FMA and its administration/scoring manual developed by Platz et al.\(^5\) into Japanese; (2) investigation of the measure's psychometric properties of the Japanese version of the FMA for the affected upper extremity in a population of patients with acute- to-chronic stages of stroke.

**Methods**

This prospective, cross-sectional, single-center study was approved by the authorized institutional human research review board at the institution governing the research. All procedures followed in this study were carried out in accordance with the institution's guidelines, and written informed consent was obtained from each patient before inclusion. All patients were recruited from the Hyogo College of Medicine Hospital in Japan from June 2016 to March 2017. This study was registered with the University Hospital Medical Information Network Clinical Trial Registry (UMIN000022188) in May 2016 as a pre-initiation condition. The study conforms to the STROBE guidelines for reporting observational studies. A completed checklist is provided in Supplemental online material.

**Participants (inclusion/exclusion criteria)**

Participant recruitment targeted adults (aged 20–100 years) with a history of ischemic or hemorrhagic stroke. The diagnosis of stroke was confirmed by a physician on the basis of review of medical records. The key inclusion criterion was the appearance of a symptom of incomplete upper extremity paresis resulting from stroke. To avoid the confounding effects of cognitive and medical conditions, the following exclusion criteria were used: (1) clear signs of dementia (A brief interview was conducted by an occupational therapist to assess the participants’ cognitive dysfunction [attention deficit, disorientation, or memory disturbance]); (2) mental disorder or aphasia that act as an obstacle to daily living; (3) excessive pain in any joint that might limit participation; and (4) severe end-stage or uncontrolled medical conditions that would interfere with participation.

**Measurement instruments**

Three instruments were used in examining the validity of the FMA: the Action Research Arm Test (ARAT), the Box-and-Block Test (BBT), and the Motor Activity Log (MAL). The FMA was commonly combined with the measures-related activity level (The WHO International Classification of Functioning, Disability and Health [ICF] definition), such as the ARAT and MAL in stroke rehabilitation studies.\(^7\) The FMA, ARAT, and MAL are ordinal scales where ceiling and floor effects could be present.\(^7,18\) Therefore, the BBT, timed grasping performance test, was also used. This instrument is usually suitable to assess the affected upper extremities in patients with mild-to-moderately hemiparesis after stroke.\(^19\)

- FMA for upper extremity (FMA)

The translated and adapted versions of the FMA motor domain (3-level ordinal scale, 33 items [4 subtests], score range, 0–66 points) and the FMA sensation domain (3-level ordinal scale, 6 items [2 subtests], score range, 0–12 points), and the FMA passive joint motion/pain domain (3-level ordinal scale, 24 items [2 subtests], score range, 0–48 points) were used. The measurement instrument has been extensively examined against other measures.\(^6,11,13,18,20–22\) It has also been used as the “gold standard” for comparing other upper extremity measures on the assessment of impairments in ICF body function and structure.\(^5\)

- Action Research Arm Test (ARAT)

The ARAT (4-level ordinal scale, 19 items [4 subtests], score range, 0–57 points) is a performance test which is representative of the major activities of the upper limb in activities of daily living. Like the FMA, it has been used as a “gold standard” on the assessment of limitations in activity (ICF definition).\(^5\)

- Box-and-Block Test (BBT)

The number of wooden blocks (2.5 cm × 2.5 cm × 2.5 cm) that can be transported from one compartment of a box to another within one minute is counted.\(^20\) The BBT is frequently used in research and rehabilitation of children and adults.\(^5\)

- Motor Activity Log (MAL)

The upper-extremity MAL-14 is a structured interview that elicits information about 14 activities of daily living.\(^23\) Patients are asked to rate how well (Quality of Movement [QOM] scale) and how much (Amount of Use [AOU] scale) they use their affected upper extremity to accomplish each ADL.

**Translation of the manual for the FMA**

To our knowledge, detailed manuals for the FMA administration and scoring were provided by Platz et al.\(^5,20\) and Sullivan et al.\(^10\) for standardization of clinical assessments. For the present study, a standardized guidebook developed by Platz et al.\(^3\) was used because their manuals were the most detailed and many photos also helped us to understand the test administration and scoring.

The translation of the FMA manual was performed according to the development of the Japanese SF-36 Health Survey\(^17\), which is a questionnaire used to measure health status in general. Transcultural adaptation is performed in four main steps: (1) an English-speaking bilingual occupational therapist performed a first translation (forward translation) from English to Japanese; (2) another English-speaking bilingual individual, who didn't know the original FMA manual, translated a first Japanese version of the FMA manual back into English (back translation); (3) a third native English-speaking bilingual individual compared the back-translation manual and the original English manual, and confirmed its appropriateness; (4) the text was adapted and modified according to his/her feedback. These four steps were repeated to the point that those two manuals (the back-translation manual and the original English manual) were checked to ensure consistency. An 11-level grading scale evaluated the consistency for each set of 63 items. The rating was from 0 (strongly disagree) to 10 (strongly agree). The transcultural adaptation ended when all 63 items rated 10 (strongly agree).
Rater training program before recruiting

The occupational therapist raters completed a training program that included: (1) practice with the guidebook for one month; (2) on-site coaching and feedback by a lead occupational therapist; and (3) data-sets from 30 patients were used for consensus meetings regarding test administration and scoring procedures.

Evaluation process

All assessments were completed in a two-day interval. Non-simultaneous reliability (test–retest reliability) was examined using two sets of data with a two-day interval for each patient. Two occupational therapist raters scored the performance of each patient independently from each other. The two raters were blind to each other’s assessments during the evaluation process. First set of each test performance of the FMA motor domain was videotaped in a prospectively standardized taping procedure for simultaneous reliability raters. Based on video information, the performance of the affected upper extremity was scored by two occupational therapists for simultaneous reliability of the FMA motor domain. The two raters were also blind to each other’s assessments during the evaluation process.

Statistical analysis

Data were analyzed using SPSS (version 22.0; IBM Corp., Armonk, NY) and the statistical package R (version 3.4.0; available at: http://www.r-project.org).

- Reliability
  The inter-rater/intra-rater reliability of the sum scores (total/subtests) were assessed using the intraclass correlation coefficient (ICC, two-way random approach with absolute agreement).
  - Agreement
    Agreement was assessed in the following three ways: (1) Bland–Altman plot (including the limits of agreement); (2) the standard error of measurement (SEM); (3) the smallest detectable change (SDC). The limits of agreement equal the mean change in scores of repeated measurements (mean_change) ± 1.96 × standard deviation of these changes (SD_change). The SEM equals the square root of the error variance of an ANOVA analysis including systematic differences (SEM_agreement). The SEM can be converted into the SDC (SDC = 1.96 × √2 × SEM), which reflects the smallest within-person change in score that, with p < 0.05, can be interpreted as a “real” change, above measurement error, in one individual (SDC_ind). Internal consistency
  - The internal consistency was expressed using Cronbach’s alpha coefficients.
  - Validity
    The convergent validity, a subtype of construct validity (how strongly a measure correlates with other related measures), was examined. The validity was determined by examining the relationships between the score of the FMA (all domains [total], motor domain, sensation domain, and joint motion/pain domain) and those of the ARAT, BBT, and MAL using the Spearman’s rho correlation coefficient. First set of each FMA assessment was analyzed for the validity.

- Floor and ceiling effects
  The score distributions were examined for floor and ceiling effects. The floor effect is the percentage of the sample scoring the minimum possible points, reflecting the extent to which scores cluster at the bottom of the scale range. The ceiling effect is the percentage of the sample scoring the maximum possible points, reflecting the extent to which scores cluster at the top of the scale range. Average scores of two sets of the FMA assessment was analyzed for the effects.

Sample size estimation

Sample size was determined using Bonett’s methods. A sample size of 28 patients was necessary for a reliability study with two raters, an ICC planning value of 0.8, a desired precision width of 0.15, and an alpha equal to 0.05.

Results

Translation of the manual for the FMA

In the first consistency match, the median of consistency score was 7.3 and the range was 3.9–9.2 (Table 1). Since the translation did not become “completely same,” the Japanese forward-translation was modified for the next phase. In the second consistency match, the median of consistency score was 9.6 and the range was 8.8–10.0 (Table 1). The second match also revealed the necessity of revision of the Japanese forward-translation. In the third consistency match, the back-translation manuals became “completely same” with the original English manuals (Table 1). The final translation scoring sheet and manuals are available in Supplemental online material.

Psychometric properties of the FMA

Since the FMA standardized manuals were successfully translated, the psychometric properties of the final translation were evaluated. The demographic and clinical characteristics of the 30 participants are shown in Table 2. No data were missing. The study sample included patients with mild-to-severe upper extremity paresis after stroke. The participants’ descriptive statistics are shown in Table 3.

Reliability

The results of the comparisons of the reliability of the sum scores of total/subtests of the FMA for the affected upper extremity (ICC) are shown in Table 4 (non-simultaneous inter-rater reliability), Table 5 (simultaneous inter-rater reliability), and Table 6 (non-simultaneous intra-rater reliability). Regarding the non-simultaneous and simultaneous inter-rater reliability, ICC ranged from 0.809 to 0.983 (95% confidence interval [CI], 0.639–0.994; p < 0.001) and 0.991–0.999 (95% CI, 0.981–0.999; p < 0.001), respectively. Regarding the simultaneous intra-rater reliability, ICC ranged from 0.994 to 0.999 (95% CI, 0.988–1.000; p < 0.001).
The reliability of the individual item scores of the outcome measures (weighted kappa) are summarized in Tables 4, 5, and 6. Regarding the non-simultaneous and simultaneous inter-rater reliability, the median of the sum scores of total/subtests of the outcome measures ranged from 0.512–0.859 to 0.951–1.000, respectively. Regarding the simultaneous intra-rater reliability, the median of the sum scores of total/subtests of the outcome measures was 1.000.
Table 5. Simultaneous inter-rater reliability and agreement.

| Fugl-Meyer Assessment | Intraclass correlation coefficient | Weighted Kappa* |
|-----------------------|-----------------------------------|----------------|
|                       | Value    | 95% CI          | P Value | Z Value | PA (%) |
| Upper extremity motor section |          |                 |         |         |        |
| A. Shoulder/elbow/forearm | 0.999    | 0.997–0.999     | p < 0.001 | 0.650 | 1.800 |
| B. Wrist               | 0.991    | 0.981–0.996     | p < 0.001 | 0.362 | 1.003 |
| C. Hand                | 0.998    | 0.996–0.999     | p < 0.001 | 0.232 | 0.644 |
| D. Coordination/speed  | 0.997    | 0.994–0.999     | p < 0.001 | 0.135 | 0.373 |
| Upper extremity motor section |          |                 |         |         |        |
| A. Shoulder/elbow/forearm | 0.995    | 0.857–1.000     | p < 0.001 | 4.643 | 97    |
| B. Wrist               | 0.960    | 0.882–1.307     | p < 0.001 | 5.986 | 97    |
| C. Hand                | 1.000    | 1.000–1.000     | p < 0.001 | 5.429 | 100   |
| D. Coordination/speed  | 1.000    | 1.000–1.000     | p < 0.001 | 5.262 | 100   |

Notes: PA, perfect agreement. SEM, the standard error of measurement. SDC, the smallest detectable change.
*Data are presented as median.

Table 6. Simultaneous intra-rater reliability and agreement.

| Fugl-Meyer assessment | Intraclass correlation coefficient | Weighted Kappa* |
|-----------------------|-----------------------------------|----------------|
|                       | Value    | 95% CI          | P Value | Z Value | PA (%) |
| Upper extremity motor section |          |                 |         |         |        |
| A. Shoulder/elbow/forearm | 0.999    | 0.998–1.000     | p < 0.001 | 0.640 | 1.773 |
| B. Wrist               | 0.998    | 0.996–0.999     | p < 0.001 | 0.451 | 1.251 |
| C. Hand                | 0.999    | 0.998–0.999     | p < 0.001 | 0.120 | 0.332 |
| D. Coordination/speed  | 0.994    | 0.988–0.997     | p < 0.001 | 0.163 | 0.453 |
| Upper extremity motor section |          |                 |         |         |        |
| A. Shoulder/elbow/forearm | 1.000    | 1.000–1.000     | p < 0.001 | 4.925 | 100   |
| B. Wrist               | 1.000    | 1.000–1.000     | p < 0.001 | 6.242 | 100   |
| C. Hand                | 1.000    | 1.000–1.000     | p < 0.001 | 5.741 | 100   |
| D. Coordination/speed  | 1.000    | 1.000–1.000     | p < 0.001 | 5.083 | 100   |

Notes: PA, perfect agreement. SEM, the standard error of measurement. SDC, the smallest detectable change.
*Data are presented as median.

Discussion

To our knowledge, this is the first study to investigate the non-simultaneous inter-rater reliability of all the FMA domains for affected upper extremities with appropriate sample size calculations based on the ICC in hemiparetic patients after stroke. In the current study, the FMA standardized manuals were successfully translated. The results indicate that the Japanese version of the FMA motor domain and all domains is highly reliable and valid for measuring upper extremity function in patients with stroke who have mild-to-severe upper extremity paresis for both non-simultaneous and simultaneous assessment.

Comparison with psychometric properties criteria

Reliability

Regarding the ICC values, Portney and Watkins suggested that reliability should exceed 0.90 to ensure reasonable validity as a general guideline for many clinical measurements. Regarding ordinal data, Landis and Koch suggested that kappa values above 80% represent excellent agreement, and those above 60% represent substantial agreement. The results of the present study (motor domain, sensation domain, and all domains) meet the ICC and kappa value criteria, indicating that the FMA with adapted standardized guidelines in hemiparetic patients after stroke is highly reliable for the motor domain, sensation domain, and all domains of the FMA. However, the interpretation of the individual item change of the FMA joint motion/pain domain in its current form for upper extremity function of stroke patients was not supported.

Agreement

For evaluative purposes, the absolute measurement error should be smaller than the minimal amount of change in the scale that is considered to be important. Although the minimal clinically
Figure 1. Bland–Altman Plot of the FMA Scores. (A) Non-simultaneous agreement of the FMA all domains. (B) Non-simultaneous agreement of the FMA motor domain. (C) Simultaneous agreement of the FMA motor domain. The horizontal lines indicate the mean difference (middle) and the upper and lower limits of agreement.
important difference on the FMA scale hasn’t been determined enough, a greater than 10% change in the FMA motor scores is considered to be a clinically meaningful improvement based on clinical experience with this scale and consultation with physical therapists and stroke neurologists. The results of the present study meet the criteria. In addition, the MDC of a measure was considered satisfactory when the MDC was less than 10% of the highest possible score on the measure. We found that only the MDC of the FMA all domains in non-simultaneous inter-rater assessment was below 10% of its highest score, indicating a satisfactory level of measurement error. On the other hand, regarding the simultaneous inter-rater assessment and the simultaneous intra-rater assessment, all the MDCs of the FMA except the FMA wrist subtest was below 10% of their corresponding highest.

**Internal consistency**

Terwee et al. proposed a criterion of 0.70–0.95 as a measure of good internal consistency. The results of the present study were higher than the criterion, indicating high correlations among the items in the scale (i.e. redundancy of one or more items).

**Validity**

A value of between 0.00 and 0.25 was considered to represent little or no relationship, a value of between 0.25 and 0.50 represented a fair relationship, a value of between 0.50 and 0.75 represented a moderate to good relationship, and value of greater than 0.75 represented a good to excellent relationship. We found that the FMA total domains and motor domain were significantly strongly correlated with the ARAT, the BBT, and the MAL. On the other hands, the FMA sensation domain and joint motion/pain domain were weakly correlated with the ARAT, the BBT, and the MAL. These results support the convergent validity of the FMA total domains and motor domain, and the discriminant validity of the FMA sensation domain and joint motion/pain domain. We believe that as far as sensation and pain are most part of the measurement domain, they do not strongly correlate to the motor instruments such as the ARAT and BBT in general.

**Floor and ceiling effects**

Floor or ceiling effects were considered to be present if >15% of the respondents achieved the lowest or highest possible score, respectively. According to the criteria, the FMA all domains and motor domain showed no floor or ceiling effects; the FMA motor subscales, sensation domain, and joint motion/pain domain showed floor or ceiling effects. The ARAT had significant ceiling effects; the BBT, MAL (AOU), and MAL(QOM) had significant floor effects. The FMA all domains and motor domain for the affected upper extremity were the only measurement instruments that did not show obvious floor or ceiling effects.

**Comparison with previous studies properties**

The Japanese version of the FMA (all domains, motor domain, and sensation domain) were found to be reliable (Tables 4, 5, and 6) with ICC values 0.90–0.99, similar to those of See et al., Sullivan et al., and Platz et al. On the other hand, the FMA joint motion/pain domain showed lower ICC values than those of Platz et al. This little difference could be due to the differences between simultaneous and non-simultaneous assessments. Validity between the FMA all domains/motor domain and motor instruments (the ARAT, BBT, and MAL) was also excellent (Table 7) and similar to previous findings. Weak correlations were observed between the FMA sensation/motion/pain domain and motor instruments (the ARAT, BBT, and MAL), which was also similar to prior findings.

**Major strength of the present study**

Extensive investigation of the psychometric properties of the FMA for the affected upper extremity after stroke has been
undertaken. However, only one study has examined inter-rater reliability of all the FMA domain with appropriate sample size based on the ICC which is the most suitable and most commonly used reliability parameter for continuous measures. The study focused on simultaneous inter-rater reliability of the FMA, therefore, the performance of the affected arm was scored, based on video information. The present study also employed all the FMA domains for the affected upper extremity with appropriate sample size calculations based on the ICC. In addition, there is no study of inter-rater reliability of all the FMA domains for affected upper extremities with appropriate sample size focusing on non-simultaneous assessment. Therefore, this is the first study to investigate the non-simultaneous inter-rater reliability of all the FMA domains for affected upper extremities with appropriate sample size calculations based on the ICC in hemiparetic patients after stroke. The results indicate that the non-simultaneous inter-rater reliability of all the FMA domains for affected upper extremities is also highly reliable for measuring upper extremity function in patients with stroke who have mild-to-severe upper extremity paresis.

In addition, the present study has four more strengths worth mentioning. First, we did the appropriate translation from English to Japanese by the recommended method. Second, regarding the translation target of evaluating functional outcome measure, a specially standardized guidebook (detailed manuals) was selected to reduce the differences among inter-raters and intra-raters for both test administration and scoring of the FMA. Third, the present study employed all the FMA domains for the affected upper extremity with appropriate sample size calculations. Many studies employed only the motor domain and/or the sensory domain, even if employed all domains, most of them have too small sample size for reliability study based on the ICC which is the most suitable and most commonly used reliability parameter for continuous measures. Fourth, in the reliability evaluation, we have mentioned even the concept of simultaneous and non-simultaneous reliability. This approach is very important for clinical trials and interpretation of the results with the FMA score.

**Differences between simultaneous and non-simultaneous assessments**

Simultaneous assessment reflects the differences between raters. On the other hand, non-simultaneous assessment reflects not only the differences between raters but also the fluctuations of patient performance. Indeed, the results of the present study showed that the differences of the non-simultaneous assessment were greater than the differences of the simultaneous assessment. Regarding the ICC, there seems to be no big difference between the two types of assessments. However, regarding the SDC, there is big difference between the assessments. Therefore, to investigate both non-simultaneous and simultaneous reliability is very important.

**Limitations**

This study did have several limitations. First, this was a single-center study. Second, we did not control for the severity of paresis or the duration of disease. Third, it will be beneficial to investigate responsiveness (i.e. standardized response mean, minimal clinically important difference, and clinically important difference) in a larger sample of patients. Future longitudinal studies involving multiple centers with stricter control are therefore warranted. Fourth, since the sample size of the present study is based on the ICC, the precision regarding the psychometric properties except reliability might decrease. Fifth, we did not explore other methods such as factor analysis or item response theory methods, for example, Rasch analysis. These approaches can show different properties such as the dimensionality of item content and the construct validity of the item structure. Sixth, although the FMA motor domain and all domains had good reliability and validity, the FMA sensation and joint motion/pain domain showed inadequate reliability, validity, and ceiling effects. This limitation must be taken into consideration by users of this tool.

**Conclusions**

The English FMA standardized manuals were successfully translated into Japanese. The results of the present study suggest that the FMA motor domain (0–66 points) and all domains (0–126 points) for affected upper extremities can reliably assess the affected upper extremities in patients with mild-to-severe hemiparesis after stroke for both non-simultaneous and simultaneous assessment.

**Contributors**

KD was Chief Investigator. SA, TT, YU, and KD provided concept/idea/project design. SA and YU provided writing. AU (Umeji), AU (Uchita), and YH provided data collection. SA, KD, TT, and YU designed the analysis. SA and YU provided data analysis. SA and TT provided project management. KD and YU provided fund procurement, facilities/equipment, and institutional liaisons. YH, KT, and YU provided administrative support. All authors contributed to data interpretation and the final version of the paper.

**Acknowledgments**

The authors would like to thank Dr Y. Miyabe for assistance with patient recruitment.

**Disclosure statement**

Mr Takebayashi serves as a consultant to Teijin Pharma Limited.

**Funding**

This work was supported by Teijin Pharma Limited (Tokyo, Japan).

**References**

1. Weinstein CJ, Stein J, Arena R, et al. Guidelines for adult stroke rehabilitation and recovery: a guideline for healthcare professionals from the American heart association/American stroke association. *Stroke*. 2016;47(6):e98–e169.
2. van der Putten JJ, Hobart JC, Freeman JA, Thompson AJ. Measuring change in disability after inpatient rehabilitation: comparison of the
responsiveness of the Barthel index and the functional independence measure. *Neurol Neurosurg Psychiatry*. 1999;66(4):480–484.

3. Baker K, Cano SJ, Playford ED. Outcome measurement in stroke: a scale selection strategy. *Stroke*. 2011;42(6):1787–1794.

4. Fugl-Meyer AR, Jaasko L, Leyman I, Olsson S, Steglin S. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. *Scand J Rehabil Med*. 1975;7(1):13–31.

5. Platz T. ARM, Arm rehabilitation measurement: manual for performance and scoring of the Fugl-Meyer test (arm section), action research Arm test, and the box-and-block test. Baden-Baden: Deutscher Wissenschafts-Verlag; 2005.

6. Gladstone DJ, Danells CJ, Black SE. The Fugl-Meyer assessment of motor recovery after stroke: a critical review of its measurement properties. *Neurorehabil Neural Repair*. 2002;16(3):232–240.

7. Santisteban L, Teremetz M, Bleton JP, Baron JC, Maier MA, Lindberg PG. Upper limb outcome measures used in stroke rehabilitation studies: a systematic literature review. *PLoS One*. 2016;11(5):e0154792.

8. Duncan PW, Propst M, Nelson SG. Reliability of the Fugl-Meyer assessment of sensorimotor recovery following cerebrovascular accident. *Phys Ther*. 1983;63(10):1606–1610.

9. Sanford J, Moreland I, Swanson LR, Stratford PW, Gowland C. Reliability of the Fugl-Meyer assessment for testing motor performance in patients following stroke. *Phys Ther*. 1993;73(7):447–454.

10. Sullivan kJ, Tilson JK, Cen SY, et al. Fugl-Meyer assessment of sensorimotor function after stroke: standardized training procedure for clinical practice and clinical trials. *Stroke*. 2011;42(2):427–432.

11. See J, Dodakian L, Chou C, et al. A standardized approach to the Fugl-Meyer assessment and its implications for clinical trials. *Neurorehabil Neural Repair*. 2013;27(8):732–741.

12. Page SJ, Fulk GD, Boyne P. Clinically important differences for the upper-extremity Fugl-Meyer scale in people with minimal to moderate impairment due to chronic stroke. *Phys Ther*. 2012;92(6):791–798.

13. Page SJ, Levine P, Hade E. Psychometric properties and administration of the wrist/hand subscales of the Fugl-Meyer assessment in minimally impaired upper extremity hemiparesis in stroke. *Arch Phys Med Rehabil*. 2012;93(12):2373–2376.e5.

14. Michaelsen SM, Rocha AS, Knabben RJ, Rodrigues LP, Fernandes CG. Translation, adaptation and inter-rater reliability of the administration manual for the Fugl-Meyer assessment. Revista Brasileira de Fisioterapia (Sao Carlos (Sao Paulo, Brazil)). 2011;15(1):80–88.

15. Lundquist CB, Maribo T. The Fugl-Meyer assessment of the upper extremity: reliability, responsiveness and validity of the Danish version. *Disability Rehabil*. 2017;39(9):934–939.

16. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*. 2000;25(24):3186–3191.

17. Fukuhara S, Bito S, Green J, Hsiao A, Kurokawa K. Translation, adaptation, and validation of the SF-36 health survey for use in Japan. *J Clin Epidemiol*. 1998;51(11):1037–1044.

18. Lin JH, Hsu MJ, Shue CF, et al. Psychometric comparisons of 4 measures for assessing upper-extremity function in people with stroke. *Phys Ther*. 2009;89(8):840–850.

19. Chen HM, Chen CC, Hsueh JP, Huang SL, Hsieh CL. Test-retest reproducibility and smallest real difference of 5 hand function tests in patients with stroke. *Neurorehabil Neural Repair*. 2009;23(5):435–440.

20. Platz T, Pinkowski C, van Wijck F, Kim IH, di Bella P, Johnson G. Reliability and validity of arm function assessment with standardized guidelines for the Fugl-Meyer Test, Action Research Arm Test and Block Test: a multicentre study. *Clin Rehabil*. 2005;19(4):404–411.

21. Rabadi MH, Rabadi FM. Comparison of the action research arm test and the Fugl-Meyer Assessment as measures of upper-extremity motor weakness after stroke. *Arch Phys Med Rehabil*. 2006;87(7):962–966.

22. Hsueh IP, Hsu MJ, Shue CF, Lee S, Hsieh CL, Lin JH. Psychometric comparisons of 2 versions of the Fugl-Meyer motor scale and 2 versions of the stroke rehabilitation assessment of movement. *Neurorehabil Neural Repair*. 2008;22(6):737–744.

23. Uswatte G, Taub E, Morris D, Vignolo M, McCulloch L. Reliability and validity of the upper-extremity motor activity log-14 for measuring real-world arm use. *Stroke*. 2005;36(11):2493–2496.

24. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropractic Med*. 2016;15(2):155–163.

25. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60(1):34–42.

26. Arwert HJ, Keizer S, Kromme CH, Vliet Vlieland TP, Meesters JJ. Validity of the Michigan hand outcomes questionnaire in patients with stroke. *Arch Phys Med Rehabil*. 2016;97(2):238–244.

27. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. *Stat Med*. 2002;21(9):1331–1335.

28. Portney L, Watkins M. *Foundations of clinical research: application to practice*. Stamford: Appleton & Lange; 1993.

29. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.

30. Smidt N, van der Windt DA, Assendelft WJ, et al. Interobserver reproducibility of the assessment of severity of complaints, grip strength, and pressure pain threshold in patients with lateral epicondylitis. *Arch Phys Med Rehabil*. 2002;83(8):1145–1150.