SYSTEMATIC REVIEW OF OUTCOME MEASURES USED IN THE EVALUATION OF ROBOT-ASSISTED UPPER LIMB EXERCISE IN STROKE

Manoj Sivan, MRCS1,2, Rory J. O’Connor, MD, FRCP1,2,3, Sophie Makower, BA(Hons), Grad. Dip. Phys1,3, Martin Levesley, BEng., PhD4 and Bipinchandra Bhakta, MD, FRCP1,2

From the 1Academic Department of Rehabilitation Medicine, University of Leeds, 2National Demonstration Centre for Rehabilitation Medicine, Leeds Teaching Hospitals NHS Trust, 3Community Rehabilitation Unit, NHS Leeds Community Healthcare and 4Department of Mechanical Engineering, University of Leeds, UK

Objective: To classify and evaluate outcome measures currently used in robot-assisted exercise trials (RAET) in stroke, and to determine selection criteria for outcome measures in future trials.

Methods: Outcome measures used in RAET were identified from MEDLINE, EMBASE, CINAHL, PubMed and PsychINFO databases. The scale items were categorized into International Classification of Functioning Disability and Health (ICF) domains. The psychometric properties of scale were rated using a standardized pro forma.

Results: Thirty outcome measures were identified from 28 published RAET. Commonly used ICF body function scales were: Fugl-Meyer (FM) (24 studies), Modified Ashworth Scale (13 studies), Medical Research Council (11 studies), Kinematic measures (8 studies) and Motor Status Score (6 studies); ICF activity scale was Functional Independence Measure (FIMTM) (9 studies); ICF participation, personal and environmental factors scales were rarely used. Standard-ized rating identified that FM, kinematic measures, Action Research Arm Test, Wolf Motor Function Test, FIMTM, and ABILHAND have adequate measurement properties for use in RAET.

Conclusion: Some of the currently used outcome measures seem appropriate for RAET. The use of the ICF framework enables selection of an appropriate combination of outcome measures depending on patient characteristics, such as severity of weakness and chronicity of stroke impairments.

Key words: rehabilitation; robotic devices; scales; ICF; psychometrics.


Correspondence address: Manoj Sivan, NIHR Academic Clinical Fellow, Academic Department of Rehabilitation Medicine, University of Leeds, D Floor, Martin Wing, Leeds General Infirmary, Leeds LS1 3EX, UK. E-mail: M.Sivan@leeds.ac.uk

Submitted August 15, 2010; accepted December 17, 2010

INTRODUCTION

Stroke is the commonest cause of severe physical disability. The annual incidence of new strokes in Europe is between 200 and 300 per 100,000 population (1). The recovery of upper limb function is generally slower and less complete than the return of mobility. Up to 85% of survivors experience some degree of paresis of the upper limb at the onset (2) and 25% report difficulty using the affected limb 5 years post-stroke (3). This is partly due to the complexity of movement required for upper limb function (4, 5). Increasing the amount and frequency of physical therapy can improve some aspects of motor recovery; however, physical therapy resource is limited. As a consequence, intervention is often inadequate in terms of the intensity and frequency required to relearn motor skills. Robot-assisted exercise can supplement conventional physical therapy.

A number of robotic devices have been developed to assist upper limb training and rehabilitation (6). Some have been evaluated in robot-assisted exercise trials (RAET) with patients after stroke. A meta-analysis of randomized controlled trials showed a significant improvement in upper limb motor function and no significant change in activities of daily living (ADL) function with upper limb robotics (7). This apparent lack of effect in real-life activities may relate to the effect size of the intervention and the outcome measures used in the studies. Larger trials are needed to confirm or refute the findings of smaller scale robot studies done so far. In designing larger and more expensive studies it is vital that appropriate outcome measures are used.

There is lack of consensus on the outcome measures that should be used in RAET. Most published clinical trials have used comparable outcome measures. The rationale for this is to allow comparison across studies and pooling of data for systematic review purposes. There has been emphasis on measuring change at the impairment level (either through kinematic assessment or impairment-based rating scales) rather than change in functional activities of daily living. Some recently published clinical studies have incorporated outcomes that reflect day-to-day activities. There is limited literature describing how to select outcome measures based on the nature of the intervention and the patient’s clinical features.

The aims of this systematic review are: (i) to identify the outcome measures that have been used in RAET, classify them using the International Classification of Functioning, Disability and Health (ICF) (8) and report on their psychometric properties; and (ii) to determine the factors that should be considered when selecting outcome measures in future trials. The domains described in the ICF conceptual framework of health condition, body functions (and structures), activities and participation,
and personal and environmental factors, although related, have no necessary causality between them, making measurement of all these domains necessary (9).

**METHODS**

The systematic review was undertaken in 3 stages:

**Stage 1. Identify clinical trials involving robot-assisted arm therapy in patients after stroke and describe the outcome measures used**

A search of MEDLINE, EMBASE, CINALH, PubMed and PsychINFO databases was performed to identify relevant RAET. The keywords used were: stroke, upper limb, arm, rehabilitation, motor, recovery, robot, computer, training, therapy, physiotherapy, function, study and trial. From the initial search, all abstracts were reviewed. The inclusion criteria for this review were: (i) study involving participants with diagnosis of a stroke; (ii) upper limb exercise assisted using a robot device. For this review, a robotic device was defined as any technology that has the ability to assist arm movement for therapeutic exercises; (iii) at least one outcome measure used in the study.

Studies of robot devices involving only healthy volunteers were excluded.

This stage was undertaken by authors MS, SM, ML and BB and their lists were cross-referenced with each other’s to ensure all relevant RAET had been identified.

**Stage 2. Classify the content of the outcome measures used in these RAET according to the three main ICF domains**

The content of each scale (identified in stage 1) was classified in terms of the ICF categories (8):

- Body functions and structures: functions refer to physiological functions of body systems including psychological. Structures are anatomical parts or regions of body and their components. Impairments are problems in body function or structure.
- Activity: activity refers to execution of a task by an individual. Limitations of a task are defined as difficulties an individual might experience in completing a given activity.
- Participation: involvement of an individual in a life situation. Restrictions to participation describe difficulties experienced by the individual in a life situation or role.

Contextual factors: these include personal and environmental factors that influence the relationship between the different components.

This stage was undertaken by authors MS, SM, ROC and BB.

**Stage 3. Describe the measurement properties of the identified outcome measures in patients after stroke**

A search of the same databases (used in Stage 1) was undertaken to identify RAET involving stroke participants and describing the properties of the identified scales from stage 1. The keywords used were: the name of the outcome measure, stroke, validity, reliability, questions, items, consistency, minimal clinically important difference (MCID), responsiveness, floor effect, ceiling effect and agreement. A measure-

<table>
<thead>
<tr>
<th>Table 1. Definition and standards for the evaluation criteria</th>
<th>Definition</th>
<th>Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reliability</strong></td>
<td>Reproducibility is the extent to which the same results are obtained on repeated administrations of the same questionnaire by same person (test re-test) or different people (inter-rater). Internal consistency assesses the homogeneity of the scale items (23).</td>
<td>Reproducibility (test-retest or inter-observer) – Intraclass correlation coefficient (ICC) or kappa value – excellent or high &gt; 0.75, moderate 0.4–0.74 and poor &lt; 0.40 (23, 24). Internal consistency – Cronbach’s α excellent &gt; 0.8, adequate 0.70–0.79 and low &lt; 0.70 (25, 26). Correlation coefficient value (r) – excellent &gt; 0.60, adequate 0.3–0.6 and poor &lt; 0.3 (25). ROC analysis – area under curve (AUC) excellent &gt; 0.9, adequate 0.7–0.9 and poor &lt; 0.7 (27).</td>
</tr>
<tr>
<td><strong>Validity</strong></td>
<td>The extent to which the scale measures what it intends to measure. Content validity is the extent to which the measure is representative of the conceptual domain. Criterion validity (concurrent, convergent, predictive) is the degree to which the measure correlates with a gold standard. For most of the functional scales, there is no gold standard and hence construct validity is used. Construct validity is determined by examining the hypothetical relationship between the measure and other similar measures (23).</td>
<td></td>
</tr>
<tr>
<td><strong>Responsiveness</strong></td>
<td>The ability of the instrument to detect changes that have occurred accurately over time (28). Minimal clinically important difference (MCID) – the smallest difference in score in the domain of interest that patients perceive as beneficial or that would be clinically meaningful. Floor and ceiling effects – the extent to which scores cluster at the bottom or top, respectively, of the scale range.</td>
<td>Change in score – the effect size is calculated by the observed change in score divided by the standard deviation of baseline score. Large &gt; 0.8, moderate 0.5–0.8 and small &lt; 0.5 (29, 30). Other methods: Standardized response mean (SRM) ROC analysis – area under curve (AUC) Statistical significance p-value Correlation values of observed change compared with change in other scales. MCID – described as a score value. Floor and ceiling effects – expressed as percentage of the number of scores clustered at bottom/top. Excellent 0%, Adequate &lt; 20%, poor &gt; 20% (25).</td>
</tr>
<tr>
<td><strong>Acceptability</strong></td>
<td>Respondent burden – is the length and content acceptable to the intended participants (participants with disability)? Administrative burden – how easy is the tool to administer, score and interpret? Cost implications?</td>
<td>Respondent burden – Excellent: brief (&lt; 15 min) and acceptable, Adequate: either longer or some problems of acceptability. Poor: both lengthy and problems of acceptability (25). Administrative burden – Excellent: scoring by hand, easy to interpret, Adequate: computer scoring, obscure interpretation, Poor: costly and complex scoring/interpretation (25).</td>
</tr>
</tbody>
</table>

*M. Sivan et al.*
Outcome measures for robot-assisted arm rehabilitation

Stage 1
A total of 28 RAET involving 16 robot devices met the inclusion criteria for this review. Table III summarizes the outcome measures used in these studies. The commonly used scores were Fugl-Meyer (FM; 24 studies), Modified Ashworth Scale (MAS; 13 studies), Medical Research Council power grading scale (MRC; 11 studies), Functional Independence Measure (FIM; 9 studies) kinematic measurements (8 studies) and Motor Status Score (MSS; 6 studies).

Stage 2
The individual items within each scale were classified into the 3 ICF domains (Table S1, available from: URL://http:jrm.medical-journals.se/article/abstract/10.2340.16501977-0674). Based on the overall item content, each scale was categorized as belonging to one of the ICF domains. Fig. 1 summarizes the classification of all the scales into impairment, activity, participation and contextual factor categories. Fifteen scales were identified as body function based outcome measures, 10 activities based, 2 participation based and 3 identified as contextual factor outcome measures.

Stage 3
Studies investigating the measurement properties of these scales in stroke participants were identified, and the evidence on different properties consolidated. The properties for each of the scales are described in this section. Tables IV and V summarize the properties of all the scales.

Table II. Outcome measure abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Outcome measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAT</td>
<td>Arm Motor Ability Test</td>
</tr>
<tr>
<td>ARAT</td>
<td>Action Research Arm Test</td>
</tr>
<tr>
<td>AS</td>
<td>Ashworth Scale</td>
</tr>
<tr>
<td>BBT</td>
<td>Box and Block Test</td>
</tr>
<tr>
<td>BI</td>
<td>Barthel Index</td>
</tr>
<tr>
<td>CAHAI</td>
<td>Chedoke Arm and Hand Activity Inventory</td>
</tr>
<tr>
<td>CMSA</td>
<td>Chedoke-McMaster Stroke Assessment</td>
</tr>
<tr>
<td>EMG</td>
<td>Electromyogram</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>EuroQol Quality of Life Scale</td>
</tr>
<tr>
<td>FAT</td>
<td>Frenchay Arm Test</td>
</tr>
<tr>
<td>FIM</td>
<td>Functional Independence Measure</td>
</tr>
<tr>
<td>FIM motor</td>
<td>Functional Independence Measure motor subscale</td>
</tr>
<tr>
<td>FM</td>
<td>Fugl-Meyer scale</td>
</tr>
<tr>
<td>FM motor</td>
<td>Fugl-Meyer motor subscale</td>
</tr>
<tr>
<td>fMRI</td>
<td>Functional Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>MAS</td>
<td>Modified Ashworth Scale</td>
</tr>
<tr>
<td>MFT</td>
<td>Manual Function Test</td>
</tr>
<tr>
<td>Motor AS</td>
<td>Motor Assessment Scale</td>
</tr>
<tr>
<td>MRC</td>
<td>Medical Research Council</td>
</tr>
<tr>
<td>MSS</td>
<td>Motor Status Score</td>
</tr>
<tr>
<td>NHPG</td>
<td>Nine-Hole Peg Test</td>
</tr>
<tr>
<td>NSA</td>
<td>Nottingham Sensory Assessment</td>
</tr>
<tr>
<td>RLAFT</td>
<td>Rancho Los Amigos Functional Test</td>
</tr>
<tr>
<td>RMA</td>
<td>Rivermead Motor Assessment</td>
</tr>
<tr>
<td>ROM</td>
<td>Range of Motion/Movement</td>
</tr>
<tr>
<td>SCT</td>
<td>Star Cancellation Test</td>
</tr>
<tr>
<td>SIS</td>
<td>Stroke Impact Scale</td>
</tr>
<tr>
<td>TUG</td>
<td>Timed Up and Go</td>
</tr>
<tr>
<td>TCT</td>
<td>Trunk Control Test</td>
</tr>
<tr>
<td>UMAQS</td>
<td>University of Maryland Arm Questionnaire for Stroke</td>
</tr>
<tr>
<td>VAS</td>
<td>Visual Analogue Scale</td>
</tr>
<tr>
<td>WMFT</td>
<td>Wolf Motor Function Test</td>
</tr>
</tbody>
</table>

Fig. 1. International Classification of Functioning, Disability and Health (ICF) categorization of scales used in robot studies. (For abbreviations, see Table II).
DISCUSSION

The impacts of stroke on the domains of the ICF are not always directly related to each other. The severity of impairment does not necessarily determine the limitation in activities and participation due to the varied interplay between these domains and the influence of contextual factors. Such differences may also be seen in relation to the effects of any intervention (e.g. changes at the impairment level do not necessarily translate into the other domains, e.g. participation). The selection of outcome measures is therefore crucial in the design of RAET and should aim to capture the changes in all the aspects of the health condition (in this case, stroke). Using the ICF to describe scale content should enable researchers to compare different scales and select the most appropriate ones for their trial. Individual scale measurement properties may allow targeting of the most appropriate scale to the study participants.

Published reports of RAET indicate that criteria to categorize the study participants are helpful. Some studies have considered FM scores of less than 20 or 25 to indicate severe impairment.
ment and more than 20 or 25 as moderate impairment (10, 11). Also, time since stroke has been used to indicate speed of recovery during rehabilitation. In acquired brain injury studies, participants are considered as being in the sub-acute stage of recovery if within 6 months since the event and chronic stage if more than 6 months (10, 12–14). Based on severity and time since stroke, we can therefore conceptualize participants in 4 categories (Fig. 2).

Severely impaired participants, particularly in the sub-acute stage of recovery, may need outcome measures with minimal floor effects to be able to discriminate between the score of individual participants. Kinematic measurement and the FM or MSS would be appropriate body function outcome measures for this group. Kinematic measurement, although more responsive than most of clinical scales, can be time-consuming and requires special equipment. The FM scale has been used in all RAET involving these participants and has been shown to be responsive (Table III). Among activity measures, the FIM motor subscale is suitable for use in this category and has been used in RAET with such participants (Table III).

The use of FIM motor subscale and Barthel Index (BI) is limited by responsiveness in RAET involving patients who have severe impairments that persist beyond 6 months. Two studies involving chronic stroke survivors did not identify change in FIM scale, although changes in the FM scale were reported (10, 15). This finding may also relate to the BI and FIM being measures of global physical ability that may be affected by many factors other than arm impairments. For this category, Action Research Arm Test (ARAT) may be limited by its floor effects, the Chedoke Arm and Hand Activity Inventory (CAHAI) scale would seem to be an appropriate activity scale, but has not been used in RAET. One non-robot rehabilitation study has showed the CAHAI to be more sensitive to change than ARAT in chronic participants (16).

Table IV. Psychometric properties of impairment scales

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>FM motor</th>
<th>MSS</th>
<th>CMSA</th>
<th>MAS</th>
<th>MRC</th>
<th>Kinematics</th>
<th>Grip strength</th>
<th>NHPT</th>
<th>BBT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time taken (min)</td>
<td>20</td>
<td>n/a</td>
<td>60</td>
<td>Varies</td>
<td>Varies</td>
<td>&lt;1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Number of items</td>
<td>33</td>
<td>29</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>Varies</td>
<td>Varies</td>
<td>Varies</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>3 point</td>
<td>6 point</td>
<td>7 point</td>
<td>6 point</td>
<td>6 point</td>
<td>Varies</td>
<td>Timed</td>
<td>Timed</td>
<td>Timed</td>
</tr>
<tr>
<td>Score range</td>
<td>0–66</td>
<td>0–82</td>
<td>6–42</td>
<td>0–5</td>
<td>0–5</td>
<td>Varies</td>
<td>Varies</td>
<td>Varies</td>
<td></td>
</tr>
<tr>
<td>Test-retest reliability</td>
<td>+++</td>
<td>+++</td>
<td>n/a</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>Inter-rater reliability</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>n/a</td>
<td>g++</td>
<td></td>
</tr>
<tr>
<td>Construct validity</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>n/a</td>
<td>+++</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>Responsiveness</td>
<td>++</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>2.9 kg</td>
<td>32 s</td>
<td>6/min</td>
<td></td>
</tr>
<tr>
<td>Floor effect</td>
<td>Adeq</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>Ceiling effect</td>
<td>Adeq</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>Burden</td>
<td>Adeq</td>
<td>Adeq</td>
<td>Poor</td>
<td>Adeq</td>
<td>Adeq</td>
<td>Adeq</td>
<td>Nil</td>
<td>Nil</td>
<td></td>
</tr>
</tbody>
</table>

Table V. Psychometric properties of activity and participation scales

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>BI</th>
<th>FIM motor</th>
<th>ARAT</th>
<th>WMFT</th>
<th>CAHAI</th>
<th>AMAT</th>
<th>RMA arm</th>
<th>FAT</th>
<th>Motor AS</th>
<th>ABILH-AND</th>
<th>SIS part</th>
<th>EQ5D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time taken (min)</td>
<td>10–15</td>
<td>20</td>
<td>10</td>
<td>10–12</td>
<td>25</td>
<td>45</td>
<td>20</td>
<td>3</td>
<td>20–30</td>
<td>n/a</td>
<td>n/a</td>
<td>2–3</td>
</tr>
<tr>
<td>Number of items</td>
<td>10</td>
<td>13</td>
<td>19</td>
<td>15</td>
<td>13</td>
<td>17</td>
<td>15</td>
<td>5</td>
<td>9</td>
<td>23</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Type</td>
<td>2–4 point</td>
<td>7 point</td>
<td>4 point</td>
<td>6 point</td>
<td>7 point</td>
<td>6 point</td>
<td>2 point</td>
<td>2 point</td>
<td>7 point</td>
<td>3 point</td>
<td>5 point</td>
<td>3 point</td>
</tr>
<tr>
<td>Score range</td>
<td>0–100</td>
<td>13–91</td>
<td>0–57</td>
<td>0–75</td>
<td>13–91</td>
<td>0–85</td>
<td>0–15</td>
<td>0–5</td>
<td>0–54</td>
<td>0–100</td>
<td>0–1</td>
<td></td>
</tr>
<tr>
<td>Test-retest reliability</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>n/a</td>
<td>n/a</td>
<td>+++</td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>Inter-rater reliability</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>+++</td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>Construct validity</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>+++</td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>Responsiveness</td>
<td>++</td>
<td>++</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCID</td>
<td>16</td>
<td>11</td>
<td>6</td>
<td>12</td>
<td>6.3</td>
<td>3</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>Floor effect</td>
<td>Poor</td>
<td>n/a</td>
<td>Poor</td>
<td>Poor</td>
<td>n/a</td>
<td>Poor</td>
<td>n/a</td>
<td>Poor</td>
<td>n/a</td>
<td>ADEq</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Ceiling effect</td>
<td>Poor</td>
<td>Adeq</td>
<td>Poor</td>
<td>Poor</td>
<td>n/a</td>
<td>Poor</td>
<td>n/a</td>
<td>Poor</td>
<td>n/a</td>
<td>ADEq</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Burden</td>
<td>Nil</td>
<td>Adeq</td>
<td>Adeq</td>
<td>Adeq</td>
<td>Adeq</td>
<td>Poor</td>
<td>Adeq</td>
<td>Nil</td>
<td>ADEq</td>
<td>Nil</td>
<td>ADEq</td>
<td>Nil</td>
</tr>
</tbody>
</table>

A: Two, 3 or 4 response options per item.

Scoring criteria as defined in Table I. For abbreviations, see Table II.

+++ : high/excellent; ++ : moderate; + : low/poor; n/o: no available evidence yet; adeq: adequate (acceptable) floor/ceiling effect/burden; poor: poor (unacceptable) floor/ceiling effect/burden; nil: minimal/no burden.
The third category is those with moderate impairments in the sub-acute stage of recovery. These studies require outcome measures with minimal ceiling effects to be able to discriminate between score changes observed in individual participants. Kinematic measurement and the FM or MSS are suitable body function outcome measures for this group. The ARAT, Wolf Motor Function Test (WMFT) (although both will be limited by their ceiling effects) and ABILHAND would be suitable activity scales. EuroQol Quality of Life Scale (EQ5D) will be a suitable participation outcome.

The final category of moderately impaired participants in the chronic stage will need outcome measures with high responsiveness to be able to capture the intervention effect. The FM scale seems to be less useful in such participants if the changes in impairment are small. Although the MSS was developed to be more responsive in therapeutic studies targeting arm impairments, we could not find evidence for this in RAET. The ARAT and WMFT may be suitable activity scales for these participants, as has been shown in RAET involving such participants (11, 17). ABILHAND might be useful in this group, as it captures real-life functional benefit, but the only RAET that included ABILHAND as an outcome measure did not observe significant change in scores (15).

Achievement of personalized goals can be used to capture change following intervention at an individual level (e.g. Goal Attainment Scale, Canadian Occupational Performance Measure) (18). These are suitable for monitoring individual persons, but are not appropriate for group analysis (19), which limits their usefulness in evaluating robot-assisted exercise in the context of a randomized control trial.

Participation should be considered as an important part of the evaluation with RAET. There are few scales that measure participation (e.g. Stroke Impact Scale (SIS) and SF36) that can be considered for inclusion. One robot trial included the SIS as an outcome, but recorded only the hand motor subscale of the SIS, which is an activity measure (20). The participation subscale of the SIS should be considered in future trials.

Economic evaluation should be considered as an important part of any large-scale clinical investigation of robot-assisted exercise. Therefore, when designing the trial it is important to include the use of a health utility measure (e.g. EQ-5D) and health resource utilization questionnaires within the context of RAET. One robot trial involving chronic participants did not observe any statistically significant improvement in EQ-5D scores, although statistically significant improvements were found in FM and FIM scores (15). It is possible that the EQ-5D may have lower responsiveness than FM and FIM. The responsiveness of EQ-5D in stroke participants is currently unknown (Table VI). Other measures that capture dependency and provide an estimate of care costs saving through reduction in dependency for physical assistance following robot-assisted exercise treatments, such as the Northwick Park Dependency Scale (NPDS), should be considered (21).
Personal and environmental factors have a huge influence on any intervention in rehabilitation. Patient and carer perceptions of robot-assisted exercise are important outcome measures to allow design iteration and gain information about satisfaction with the delivery of robot-assisted therapy, which could relate to the look and feel of the system (13).

The other important factor to be considered when selecting outcome measures is whether the intervention is aimed at proximal or distal upper limb muscle groups. The 3 hand function tests (Grip strength, Nine-Hole Peg Test (NHPT) and Box and Block Test (BBT)) are quick to administer and may be suitable for studies where the intervention is directed at distal limb and hand function. The hand-based robot Hand Wrist Assistive Rehabilitation Device (HWARD) trial showed greater increase in BBT and NHPT scores when compared with proximal shoulder and elbow FM and ARAT (20).

The use of functional magnetic resonance imaging (fMRI) is at an experimental stage (11) and needs to be evaluated further in robot trials. It may provide interesting insights into the recovery process, but has limitations in terms of cost and feasibility. One aspect of recovery that is neglected in RAET is measuring changes in perceptual (sensory) function arising as a result of robot-assisted arm exercise. Perceptual function is a vital part of normal movement, and evidence suggests that recovery of functional motor ability is dependent on intact sensation, spatial awareness and attention. Interactive robot-assisted exercise may improve perceptual deficits or potentially confound the benefits that might be identified in RAET. Only one RAET used a sensory assessment tool as one of the outcome measures (22). The extent of sensory impairment did not seem to influence the overall benefit from robot-assisted therapy in this study. Changes in the perceptual function need to be further researched in RAET.

Fig. 2 describes our proposed algorithm for selecting outcome measures based on the type of participants recruited to RAET. Apart from the factors mentioned in the algorithm, other factors that should be considered in selecting outcome measures are the type of assistance that the robot provides (proximal, distal or both) and available resources (e.g. to undertake measurements in terms of research staff cost and participant time). We propose that at least 4 suitable outcome measures covering the different domains of ICF could be considered as essential to understand the effects of robot-assisted exercise on arm impairments in people with stroke.

The main limitation of this review is that we have analysed in detail only those outcome measures that have been used in RAET so far. This does not necessarily mean that outcome measures not used in the trials are not suitable for use in future trials. However, this review provides a system for the selection of outcome measures, which should enable researchers to apply these criteria to the outcome they wish to explore in future trials. The future of robot-assisted rehabilitation after stroke is influenced by accurate analysis and interpretation of the observed effects. This could be accomplished by using the most appropriate outcome measures.

In conclusion, we are proposing an approach to assist researchers in selecting outcome measures in the design of future clinical trials of robot-assisted rehabilitation. We feel that a basket of outcome measures covering all domains of ICF is crucial, as it is important to measure change in each domain. The selection of outcome measures should also be based on the focus of the intervention, severity of arm impairments, time since stroke, and psychometric properties of the scales.

REFERENCES


61. Shelton FD, Volpe BT, Reding M. Motor impairment as a predictor...


77. Medical Research Council. Aids to the investigation of peripheral nerve injuries. London: Medical Research Council, Committee NI; 1943.


