Original Article

RATE

Journal of Rehabilitation and Assistive Technologies Engineering
Volume 5: 1–12

DOI: 10.1177/2055668318788036
journals.sagepub.com/home/jrt

Smallest real differences for robotic measures of upper extremity function after stroke: Implications for tracking recovery

José Zariffa[1,2,3,4], Matthew Myers[5], Marge Coahran[1] and Rosalie H Wang[1,3,6]

Abstract

Introduction: Measurements from upper limb rehabilitation robots could guide therapy progression, if a robotic assessment's measurement error was small enough to detect changes occurring on a time scale of a few days. To guide this determination, this study evaluated the smallest real differences of robotic measures, and of clinical outcome assessments predicted from these measures.

Methods: A total of nine older chronic stroke survivors took part in 12-week study with an upper-limb end-effector robot. Fourteen robotic measures were extracted, and used to predict Fugl-Meyer Assessment-Upper Extremity (FMA-UE) and Action Research Arm Test (ARAT) scores using multilinear regression. Smallest real differences and intraclass correlation coefficients were computed for the robotic measures and predicted clinical outcomes, using data from seven baseline sessions.

Results: Smallest real differences of robotic measures ranged from 8.8% to 26.9% of the available range. Smallest real differences of predicted clinical assessments varied widely depending on the regression model (1.3 to 36.2 for FMA-UE, 1.8 to 59.7 for ARAT), and were not strongly related to a model's predictive performance or to the smallest real differences of the model inputs. Models with acceptable predictive performance as well as low smallest real differences were identified.

Conclusions: Smallest real difference evaluations suggest that using robotic assessments to guide therapy progression is feasible.

Keywords

Rehabilitation robotics, upper limb, functional assessment, stroke, smallest real difference, reliability

Date received: 12 March 2018; accepted: 14 June 2018

Introduction

Robotic rehabilitation devices are most often intended to improve rehabilitation outcomes. They attempt to accomplish this goal by making it possible to deliver higher doses of therapy, while ensuring that the exercises being carried out are guided by the best available evidence on motor learning and neuroplasticity after neurological injuries.[1,2] In addition to these therapeutic applications, however, the use of rehabilitation robots as measurement tools to quantify motor function has become an area of increasing focus.[3] For example, robot-derived measures have been related to a number of clinical assessments in stroke survivors,

[1]Toronto Rehabilitation Institute – University Health Network, Toronto, Ontario, Canada
[2]Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada
[3]Rehabilitation Sciences Institute, University of Toronto, Toronto, Ontario, Canada
[4]Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada
[5]Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario, Canada
[6]Department of Occupational Science and Occupational Therapy, University of Toronto, Toronto, Ontario, Canada

Corresponding author:
José Zariffa, Toronto Rehabilitation Institute – University Health Network, 550 University Ave., #12-102 Toronto, Ontario M5G 2A2, Canada.
Email: jose.zariffa@utoronto.ca

including the Fugl-Meyer Assessment,[4–9] the Action Research Arm Test (ARAT),[8] the Functional Independence Measure,[10,11] the Motor Power Score,[6,12] and the Motor Status Score.[6,7] The models used to establish these relationships have included linear regression or correlation analyses using individual metrics,[4,6,8,13] multiple linear regression approaches[5,7,12,14] as well as nonlinear models.[10,15]

A limitation of manual clinical assessments is that they require time and trained personnel to perform, limiting the frequency of administration. In contrast, if a robotic device is being used for therapy multiple times per week and can quantify function, more frequent data points could be obtained. A natural question is then whether these frequent measurements can be used to guide day-to-day adjustments in the therapeutic plan. In this manner, training could be progressed in a data-driven and individualized manner, potentially optimizing the use of the available training time and improving outcomes. In order for the robotic measurements to be used in this way, they must be able to detect very small changes, so that the gradual changes that occur on a time scale of days instead of weeks or months can be meaningfully analyzed. To date, studies have focused primarily on the ability of rehabilitation robots to estimate function at isolated time points (validity),[4–10,12,13,16] or to examine changes on the scale of several weeks (responsiveness or recovery profiles).[11,14,15,17,18] The ability of the robotic measures to detect fine changes has received much less attention.

The smallest real difference (SRD), also known as the minimum detectable difference (MDD), can be used to quantify the smallest change that a method of measurement can reliably detect, given the expected variability or error in the measurements.[19] In contrast to test-retest reliability measures such as the intraclass correlation coefficient (ICC), the SRD can provide a practical guideline on how to interpret scores obtained on two different days. For example, if the scores obtained from a rehabilitation robot on two different days differ by an amount greater than the SRD, that change could be deemed to contain meaningful information about the patient's progression. From that perspective, if a robot-derived measure's SRD is greater than the changes that could be expected clinically over a short timespan (e.g. a few days), then we could conclude that basing day-to-day therapy adjustments on these scores is not appropriate. Conversely, if a robot-derived measure has a very low SRD, it would be of interest to understand how these frequent measurements could be incorporated into therapy planning.

We have evaluated SRD values for robot-based assessments, based on a retrospective analysis of data collected during an interventional clinical study in a population of older adult stroke survivors, using an end-effector upper limb rehabilitation robot. We further sought to understand the relationship between a regression model's SRD and its performance in predicting clinical outcome measures. Our focus here is not on introducing new robot-based assessments, but rather on evaluating the SRDs of techniques that have previously been used in the literature.

Methods

Study participants

Nine older adult stroke survivors took part in this study. Inclusion criteria were to be 60 years old or older, at least 6 months post stroke, have completed all outpatient stroke rehabilitation, to have an upper limb recovery between stages 3 to 5 (out of 7) for the arm on the Chedoke McMaster Stroke Assessment (CMSA) Stages of Motor Impairment,[20] and to be able to attend up to four visits per week at the clinic. Participants were excluded if they had significant upper limb neurological or musculoskeletal conditions other than stroke, or shoulder subluxation or significant pain that limited active mobility treatment. Demographic and stroke history information are provided in Table 1.

Additionally, six healthy volunteers without a history of stroke (40.8+/−15.7 years, two males), each took part in a single session with the robot in order to provide normative data that could be used to characterize the range of expected values for each robot-derived measure (see methods below).

Data collection

This study used data from an interventional study whose outcomes will be reported elsewhere. In brief, it was a pilot study designed as a multiple single-subject research study. The study consisted of a baseline phase, an intervention phase, and a second baseline phase (i.e. intervention withdrawal or maintenance phase). The intervention phase consisted of eight weeks of training with an upper limb robotic rehabilitation device in an outpatient clinic, in a therapy program that combined therapy goal setting and "homework." The robot used was an end-effector table-top robot, described in Lu et al.[21] and Huq et al.[22,23] and shown in Figure 1(a). Figure 1(b) provides a timeline of the sessions involved in the study and shows which time points were used for the analysis presented here.

The outcome measures of upper limb function used in the analysis presented here were the motor component of the Fugl-Meyer Assessment – Upper Extremity (FMA-UE)[24] and the ARAT.[25]

All sessions with the robot began with a calibration phase lasting approximately 10 min and consisting of

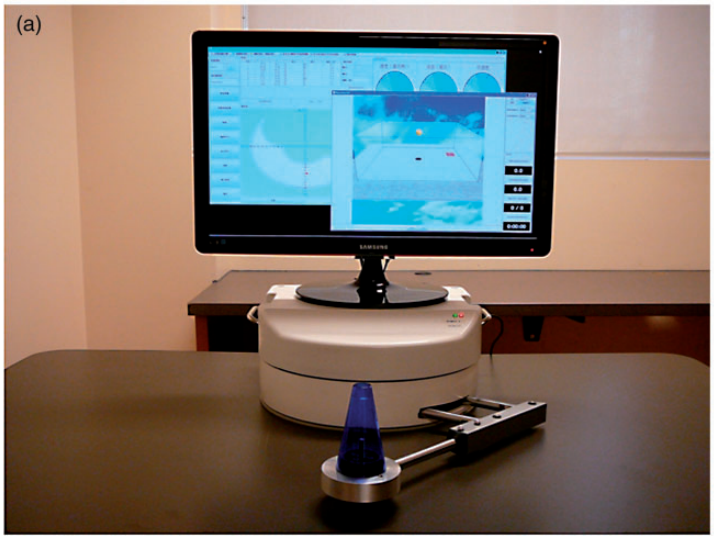
Table 1. Participant demographics and stroke history.

Participant number	Age at enrollment (years)	Sex	Affected side	CMSA stage	Hand dominance	Time post-stroke (at enrollment, years and months)	Type of stroke	Relevant medical history
1	63	F	R	3 (arm) 2 (hand)	R (pre-stroke) L (now, for most tasks)	10 years	L ischemic stroke with secondary extension	Occasional osteoarthritis joint pain (knee), cataract surgeries both eyes, anxiety/depression managed with medications
2	68	M	L	3 (arm) 2 (hand)	R	20 years	L ischemic stroke, basilar artery	Myocardial infarct
3	62	M	L	3 (arm) 3 (hand)	R	9 months	R hemorrhagic stroke, near R based ganglia extending into external capsule and corona radiata	None relevant
4	73	M	L	3 (arm) 4 (hand)	L (pre-stroke)	5 years, 1 month	R ischemic stroke, motor area	Atrial fibrillation; hypotension; fatigue; hyperlipidemia; hypothyroidism
5	60	M	L	3 (arm) 4 (hand)	L (pre-stroke) R (now)	10 years, 10 months	R ischemic stroke, location unknown	Atrial fibrillation, diabetes
6	65	M	L	3 (arm) 3 (hand)	R	2 years	R hemorrhagic stroke, basal ganglia and lentiform	Hypocholesterolemia, hypothyroidism
7	67	F	L	3 (hand) 3 (arm)	R	23 years	Scans did not indicate type of stroke; multiple area cortex involvement	Lateral epicondylitis in R, anxiety managed with medications
8	72	M	R	3 (hand) 3 (arm)	R	10 months	L hemorrhagic stroke	Hypercholestrolemia; hypertension; diabetes
9	65	F	L	3 (hand) 4 (arm)	R	1st stroke 14 years 2nd stroke 4 years	Scans revealed 100% occlusion of carotid artery	None relevant

CMSA: Chedoke McMaster Stoke Assessment.
Note: Scale 1 = flaccid paralysis to 7 = normal.

the following tasks: (1) Active range of motion: The participant was asked to move the robot arm around the workspace, tracing out the largest area they could reach on their own using their affected upper extremity. They were reminded by a research therapist or therapy assistant not to use compensatory motions like leaning forward or rotating at the trunk to complete the task. (2) Forward reaching: A sequence of 10 targets was presented to the participant. These targets alternate in location between directly forward (toward the robot) and directly backward (toward the patient) in the horizontal workspace from the start position. The targets were all located at the edge of the participant's active range of motion in the given direction, as recorded from the active range of motion task. This forward motion task was performed first without any resistance from the robot, then again with the robot exerting a damping (resistive) force against the direction of motion. The resistance gradually decreased as the task progressed. (3) Passive range of motion: The therapist used hand over hand guidance on the participant's hand to demonstrate the range of motion of the participant's upper extremity in the horizontal workspace. Again, care was taken to avoid compensatory movement.

During the intervention phase of the study, the calibration phase was followed by target exercises, configured by the therapist specifically for each participant, their abilities, and their movement goals. A sequence of targets was placed within the participant's range of motion, and haptic feedback (assistance or resistance) was optionally applied for each target. The participant's task was to reach each target in succession, using the affected upper extremity. As the participant gains strength, stability, and range of motion, the therapist altered the amount of haptic feedback given and the location of targets within the workspace to progressively extend the range of motion. After the target



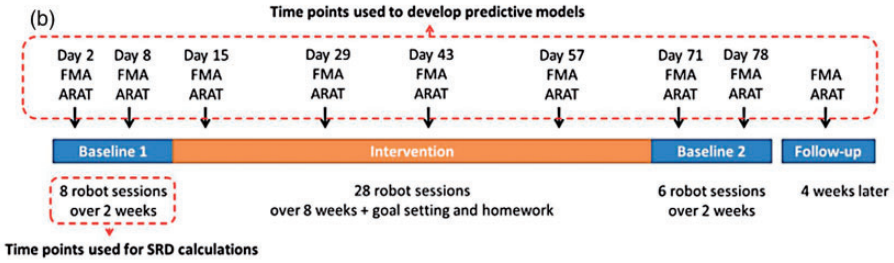


Figure 1. (a) The end-effector robot used in this study. (b) Timeline of the interventional study from which the data for this analysis were drawn. The predictive models were developed using nine time points at which both clinical and robotic measures were available (top red dotted box). The SRDs of the individual robotic measures as well as of the predictive model outputs were computed based on the initial baseline phase, which provided repeated measures during a period in which the participants were expected to be stable (bottom red dotted box). Note that the robot sessions during the baseline periods were for assessment only, while the robot sessions during the intervention period included both treatment and assessment.

exercises, participants also performed game-based exercises for part of each session. The game-based exercises did not include targets that could be used to segment the upper extremity movements.

All study procedures were approved by the Institutional Review Board of the University Health Network in Toronto, Ontario, Canada.

Robotic measures

A total of 14 measures were extracted from the robot sensor data: 8 from kinematics and 6 from range-of-motion (ROM). As this was a retrospective analysis of data from an interventional study, we focused on measures that can be extracted from therapy sessions, rather than requiring dedicated assessment procedures.[26]

For kinematic information, only portions of the session where the robot was not providing any haptic force (assistance or resistance) were used. Most of these measurements were taken from a calibration exercise at the beginning of the session (forward reaching without resistance, responsible for tailoring the subsequent exercises to the speed and smoothness of user movement), because it was always performed with no force applied. However, forceless segments from any target-based exercise were used. Discrete movements were extracted, using the time between the appearance of a new target on the display and the successful entry of the end effector into the target area, as recorded by the robot. Movements with fewer than 10 samples were discarded (using a sampling rate of 100 Hz). The game-based exercises were not used in the analysis, because they did not include targets that could be used for movement segmentation.

Eight smoothness and velocity-based measures, drawn from previous literature,[15,26,27] were extracted from each movement and averaged over each session. The velocity metrics were mean velocity and peak velocity. The smoothness measures were root mean square (RMS) jerk (normalized by movement duration), mean

rectified jerk (normalized by peak velocity), number of peaks (normalized by movement duration), path smoothness (shortest path length divided by actual path length), speed smoothness (mean velocity divided by peak velocity), and modified spectral arc length (SPARC[27]).

We additionally incorporated information from the active and passive ROM assessments that were performed during the calibration phase at the beginning of each session. Six metrics were extracted from the ROM data: X range, Y range, and total reachable area, for each of passive and active ROM.

The definitions of the measures are listed in Table 2.

Predictive models

In addition to evaluating the SRDs of individual robotic measures, it is necessary to evaluate the SRDs of the clinical estimates that can be derived from them. The robotic measures described in the previous section were used to construct predictive models to estimate the concurrent value of two outcome measures: the motor component of the FMA-UE and the ARAT. The models were constructed using multiple linear regression.

We are interested in understanding how the SRDs of the predicted scores relate to two factors: (1) the prediction accuracy of the model, and (2) the SRDs of the inputs to the model. In order to answer these questions, it is beneficial to examine multiple different models with varying predictive accuracies and input SRDs. To generate an appropriate collection of models, instead of conducting a variable selection process for the multiple linear regression (e.g. a stepwise regression), we constructed a model from every possible combination of

Table 2. Description of robotic measures.

Measure	Description	Note
Mean velocity[15]	$v_{mean} = \frac{1}{N}\sum_{n=1}^{N} v_n$	
Peak velocity[15]	$v_{peak} = \max_{movement}(\boldsymbol{v})$	
RMS jerk[15]	$J_{RMS} = \frac{RMS\left(\frac{d^2}{dt^2}\boldsymbol{v}\right)}{t_{movement}}$	Normalized by movement duration
Mean-rectified jerk[15]	$J_{MR} = \frac{\frac{1}{N-2}\sum_{1}^{N-2}\left\lvert\frac{d^2}{dt^2}\boldsymbol{v}\right\rvert}{v_{peak}}$	Normalized by peak velocity
Number of peaks[6,26]	$N_{peaks} = \left\lvert\left\{v_n > v_{n+1} \cap v_n > v_{n-1}\right\}_{n=2}^{N-1}\right\rvert$	
Path smoothness[6,15]	$S_{path} = \frac{\sqrt{(x_N-x_1)^2+(y_N-y_1)^2}}{\sum_{i=1}^{N-1}\sqrt{(x_{i+1}-x_i)^2+(y_{i+1}-y_i)^2}}$	Shortest path length divided by actual path length
Speed smoothness[15]	$S_{speed} = \frac{v_{mean}}{v_{peak}}$	
SPARC[27]	$SPARC = -\int_0^{\omega_c}\sqrt{\left(\frac{1}{\omega_c}\right)^2+\left(\frac{d\hat{V}(\omega)}{d\omega}\right)^2}\,d\omega$ Where: $\hat{V}(\omega) = \frac{V(\omega)}{V(\omega=0)}, V(\omega) = FFT(\boldsymbol{v}(t))$ $\omega_c = \min\left\{\omega_c^{max}, \min\left\{\omega \lvert \hat{V}(r) < \bar{V}\forall r > \omega\right\}\right\}$ $\bar{V} = 0.05, \omega_c^{max} = 40\pi s^{-1}$	The arc length of the FFT of the velocity, from 0 to an adaptive cut off frequency. ω_c^{max} is the maximum cut off frequency, and $\bar{V}$ is an amplitude threshold.
Passive ROM X	$ROM_{X,Passive} = \max_{\substack{Passive\\Exercise}}(\boldsymbol{x}) - \min_{\substack{Passive\\Exercise}}(\boldsymbol{x})$	
Passive ROM Y	$ROM_{Y,Passive} = \max_{\substack{Passive\\Exercise}}(\boldsymbol{y}) - \min_{\substack{Passive\\Exercise}}(\boldsymbol{y})$	
Passive ROM Area	$ROM_{Area,Passive} = \frac{1}{2}\sum_{n=1}^{N-1}\left(det\begin{bmatrix}x_n & x_{n+1}\\ y_n & y_{n+1}\end{bmatrix}\right) + det\begin{bmatrix}x_N & x_1\\ y_N & y_1\end{bmatrix}$	Polygon area
Active ROM X[26]	$ROM_{X,Active} = \max_{\substack{Active\\Exercise}}(\boldsymbol{x}) - \min_{\substack{Active\\Exercise}}(\boldsymbol{x})$	
Active ROM Y[26]	$ROM_{Y,Active} = \max_{\substack{Active\\Exercise}}(\boldsymbol{y}) - \min_{\substack{Active\\Exercise}}(\boldsymbol{y})$	
Active ROM area	$ROM_{Area,Active} = \frac{1}{2}\sum_{n=1}^{N-1}\left(det\begin{bmatrix}x_n & x_{n+1}\\ y_n & y_{n+1}\end{bmatrix}\right) + det\begin{bmatrix}x_N & x_1\\ y_N & y_1\end{bmatrix}$	Polygon area

N: number of samples in a given movement; $\boldsymbol{v}$: velocity values at each time sample; v_n: velocity at sample n; $\boldsymbol{x}$,$\boldsymbol{y}$: Cartesian coordinates of the end effector at each time sample; x_n,y_n: Cartesian coordinates of the end effector at sample n; $|\cdot|$: cardinality of a set.

the 14 input variables (including all models with numbers of variables between 1 and 14). This resulted in a total of 16,383 models, which can be expected to cover a range of accuracies and input SRDs.

In order to train the models, the robotic measures had to be mapped to the clinical assessments according to time of occurrence. For each of the nine assessment time points (Figure 1(b)) per stroke survivor participant, all sessions within four days of the assessment day were identified, and the robotic measures averaged over those days to obtain the predictive model inputs. One participant performed 11 assessments rather than 9, 1 performed 8, and over the entire dataset 6 of the assessments were removed from analysis due to missing calibration data, resulting in 76 assessments for use in regression modeling. Prior to constructing the models, each of the robotic measures was divided by the standard deviation over all time points from all participants for that measure, and thus normalized to have a variance of 1. The models were trained and evaluated using a leave-one-subject-out cross-validation process. Note that because of the cross-validation process, each participant may have slightly different coefficients for a given model. Each model is therefore identified by the input variables used, rather than by the coefficients.

Statistical analysis

SRD computations were performed according to Beckerman et al.[19] In brief, a two-way analysis of variance (ANOVA) was conducted with the participant modeled as a random variable and the day as a fixed variable. The within-participant variance was obtained from the results of the ANOVA. Then, the SRD was computed according to equation (1).

$$SRD = 1.96 \times \sqrt{2} \times \sqrt{within - participant\ variance} \quad (1)$$

SRD calculations for the robotic measures were carried out using seven baseline measurements obtained on different days before the training began (most participants had eight baseline measurements, as per Figure 1(b), but seven were used here due to some participants missing data points). Before computing the SRD, each of the metrics was normalized using the maximum absolute value observed for that metric across all stroke survivors and able-bodied participants. This normalization was designed to make the SRD values easier to interpret, by expressing them as a fraction of the estimated maximum value for each measure. The inclusion of data from the healthy volunteers in the normalization calculation ensured that the estimated maximum was not artificially restricted by the impairment of the participants.

Similarly, the SRDs for the predicted FMA-UE and ARAT scores were computed using the values predicted from the robot measures on each of the seven baseline days. In this manner, one SRD was obtained for each of the two outcome measures, for each of the 16,383 models described in the previous section.

Once the SRDs of the predicted FMA-UE and ARAT values were obtained for all of the predictive models, two relationships were investigated and quantified by means of the coefficient of determination (R^2) of a linear regression:

- SRD of a model's output vs. predictive accuracy of that model. The predictive accuracy was measured as the coefficient of determination between the correct and estimated values obtained during the leave-one-subject-out cross-validation process.
- SRD of a model's output vs. the SRDs of its input variables. The normalized input SRDs were weighted by their absolute coefficients in the regression model, normalized by the sum of the absolute coefficients (excluding the constant term), to arrive at a single value representing the combined SRDs of the inputs.

Lastly, the ICC for each of the robotic measures and predicted FMA-UE and ARAT scores was computed again using the seven baseline measurements.[28] The ICC is a measure of reliability and can provide insight into what benefits may be gained from using the SRD. The ICC results were analyzed using the same methods as the SRD analysis above.

Results

Robotic measures

The normalized SRDs of the robotic measures are shown in Figure 2(a). The values obtained ranged from 0.088 (for AROM Y) to 0.269 (for RMSJerk), meaning that the SRDs range from 8.8% to 26.9% of the maximum value observed in the dataset. For comparison, the ICCs of these measures are shown in Figure 2(b) and were found to range from 0.884 (for SPARC) to 0.989 (for AROM Y). This comparison suggests that high ICCs do not necessarily imply low SRDs.

Predicted scores on outcome measures

Figure 3 provides a visualization of the SRDs observed for the outputs of the 16,383 regression models constructed for each of FMA-UE and ARAT. The SRDs ranged from 1.3 to 36.2 for the FMA-UE and from 1.8 to 59.7 for the ARAT. Figure 3 additionally shows to

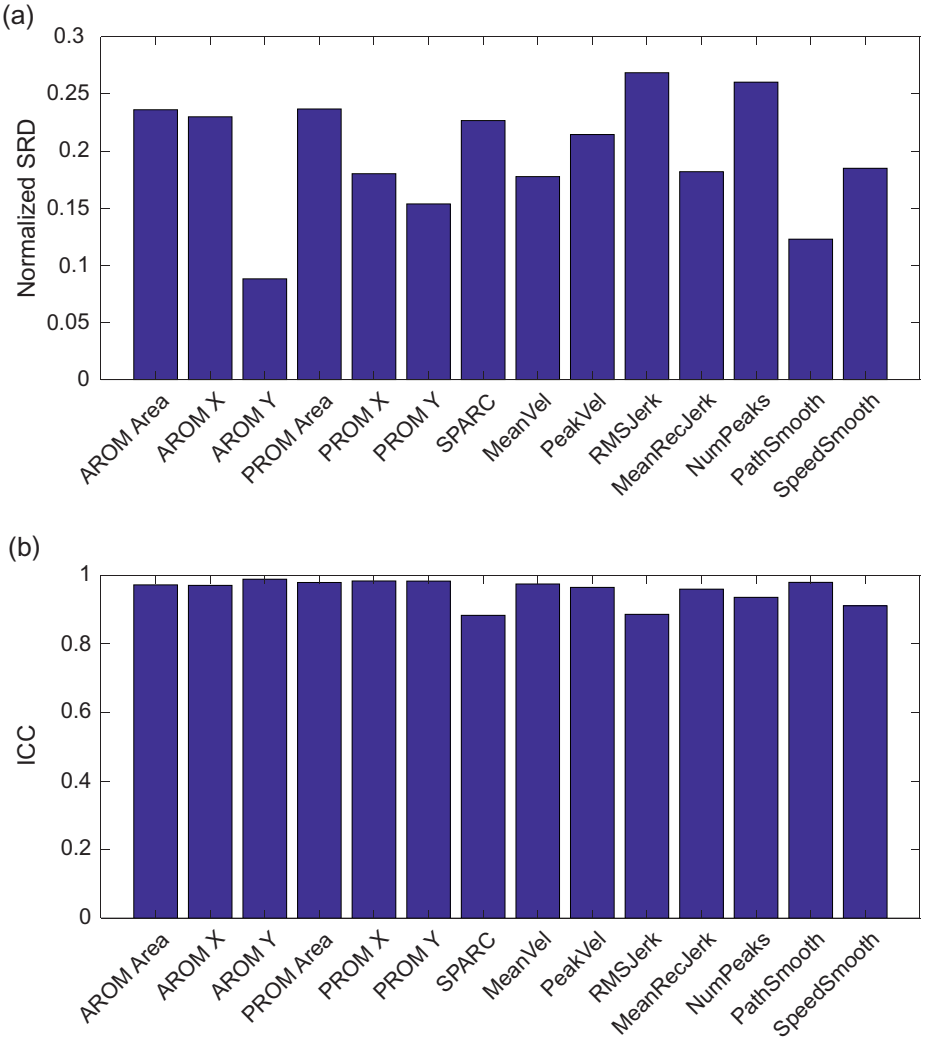


Figure 2. (a) SRD values measured for each of the robotic measures. Values are normalized to the maximum value observed for each metric across all stroke survivor participants and six healthy participants used for normative data. (b) ICC values measured for each of the robotic measures. SRD: smallest real difference; ICC: intraclass correlation coefficient.

what extent the SRD of a model's output was related to the predictive performance of that model, or to the weighted SRDs of its inputs. Although all of these relationships were statistically significant because of the large number of data points, the coefficients of determination were generally extremely low and few trends could be observed. The only exception was the relationship between model R^2 and output SRD for the ARAT, for which a coefficient of determination of 0.179 was found.

Figure 4 provides the analogous information for the ICC results. The ICCs ranged from −0.04 to 0.98 for the FMA-UE and from −1.07 to 0.981 for the ARAT, but with high values occurring much more frequently. Once again, however, no relationships could be found between the ICC value and a model's predictive performance or with the weighted ICCs of its inputs.

Trade-off between accuracy and SRD

As revealed by Figure 3, selecting the model with the highest R^2 value may not necessarily yield the smallest SRD. The question therefore arises of whether a trade-off should be sought between these two metrics. It may be possible to find a model with only slightly worse predictive performance, but a much more advantageous SRD. Table 3 illustrates several such possible trade-offs. For example, it appears that one possibly advantageous strategy may be to set a minimum acceptable threshold for the R^2, and then find the model that minimizes the SRD while still satisfying the constraint on the R^2. Using this strategy for the FMA-UE, an SRD of 1.4621 (on a 66-point scale) can be achieved in combination with an R^2 of 0.4390, and for the ARAT, an SRD of 2.6803 (on a 57-point scale) with an R^2 of 0.4246.

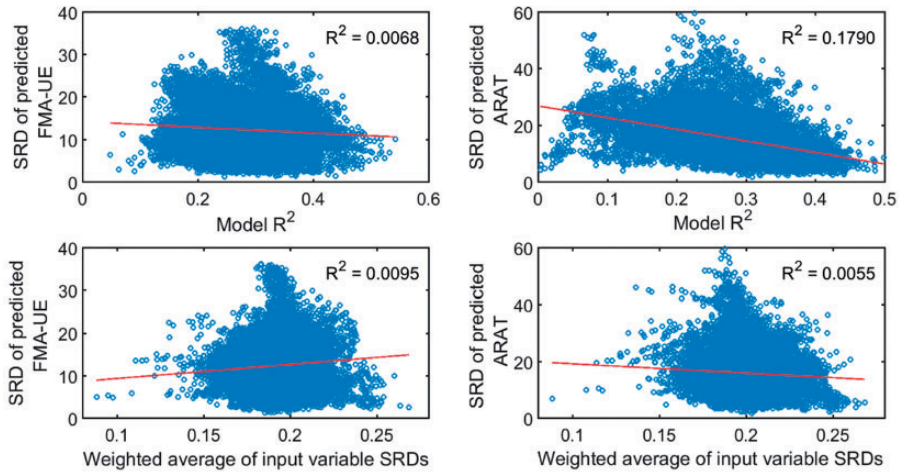


Figure 3. Top row: Relationships between the predictive performance (R^2) of each model and the SRD of its outputs, for the FMA-UE (left) and the ARAT (right). The red lines show the results of a linear regression for these two variables. Bottom row: Relationships between the weighted SRDs of each model's inputs and the SRD of its outputs, for the FMA-UE (left) and the ARAT (right). Note that the SRDs should be interpreted in the context of a 66-point scale for the FMA-UE and a 57-point scale for the ARAT. FMA-UE: Fugl-Meyer Assessment – Upper Extremity; ARAT: Action Research Arm Test; SRD: smallest real difference.

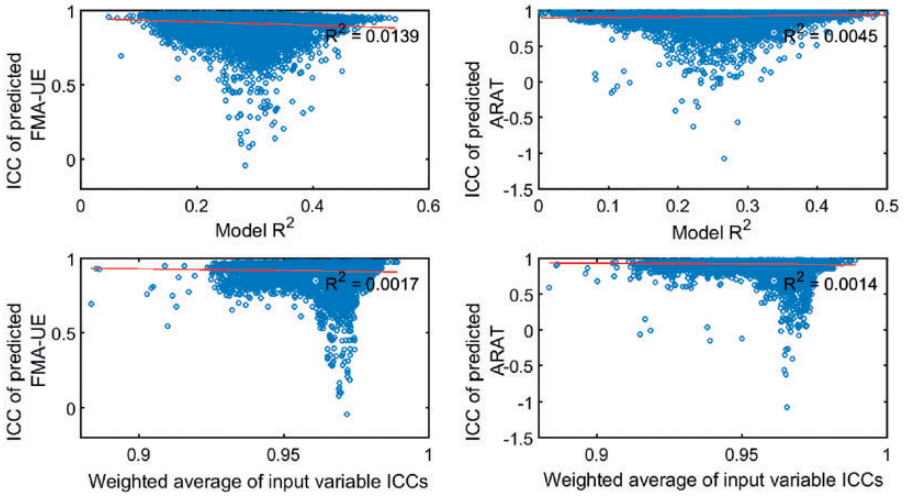


Figure 4. Top row: Relationships between the predictive performance (R^2) of each model and the ICC of its outputs, for the FMA-UE (left) and the ARAT (right). The red lines show the results of a linear regression for these two variables. Bottom row: Relationships between the weighted ICCs of each model's inputs and the ICC of its outputs, for the FMA-UE (left) and the ARAT (right). FMA-UE: Fugl-Meyer Assessment – Upper Extremity; ARAT: Action Research Arm Test; ICC: intraclass correlation coefficient.

Discussion

Upper limb rehabilitation robots have the potential to be used as measurement devices capable of evaluating the user's motor function. The information obtained in this manner could conceptually be used to track recovery with a fine temporal resolution, and guide adjustments to a therapeutic plan, such as the progression of the type and difficulty of exercises. This application would, however, require that the measurement error of the robotic metrics be comparable or smaller to the functional changes that can be expected on the time scale of a few days. This study used the SRD to gain insight into the plausibility of using rehabilitation robots in this manner. The SRD is based on the within-individual variance in repeated measurements, and therefore reflects the reliability of the measure. It additionally relates to responsiveness by showing whether a change over time reflects an underlying recovery process. Given our interest here in guiding therapy progression, the SRD has the advantage of being easily interpretable by providing a direct point

of comparison for an observed change. For a more in-depth discussion of the relationships between reliability, responsiveness and the SRD, the reader is referred to Beckerman et al.[19]

Raw robotic measures were found to have relatively high SRDs in this data (ranging from approximately 10% to 25% of the available range). On the other hand, FMA-UE and ARAT scores predicted from the robotic measures had a very wide range of SRDs, depending on the regression model used, and the SRD of a given model was not easily predicted by either the predictive accuracy of that model or the SRDs of its input variables. It was observed that explicitly seeking a trade-off between SRD and predictive accuracy resulted in models with an acceptable balance.

From a methodological standpoint, the SRD is a clinically intuitive means to quantify the expected day-to-day variations. ICC values were generally found to be quite high. This comparison shows that most models provided results across different baseline days that were certainly correlated, but that in many cases there was still too much variability in the daily measurements to use them as the basis for any short-term clinical decision-making. The SRD provides a more interpretable means to determine which models could support this type of application.

Interpreting the results presented here requires some estimate of how much change might occur over a time scale of a few days, which is different from how outcome measures are typically used, with time points separated by several weeks. Although differences in user function cannot be expected to change widely from one day to the next, a majority of outcome measures (including the FMA-UE and ARAT) use ordinal scales. There will, therefore, be a specific day when the user first saw an increase in score. In other words, there will be a specific day when the underlying continuous process of recovery translated into a discrete step in the ordinal measure being used. As such, as a general approximation independent of the specific measure, there will be days when an increase of one to two points can be expected. From Table 3, we can see that by seeking a trade-off between the R^2 and SRD of the predictive models, SRD values under 2 for the FMA-UE and under 3 for the ARAT could be achieved. These results are within or close to the target range for these models to be usable to track day-to-day changes.

The predictive performance (R^2) of the multilinear regression models obtained here was in the range of 0.5–0.55 in the best case, and in the range of 0.4–0.45 for the trade-off models. This performance is in line

Table 3. Examples of possible trade-offs between SRD and R^2 model metrics.

Objective	SRD	R^2	Input variables of selected model
FMA-UE			
Maximize R^2	10.1695	0.5428	AROM area, AROM X range, AROM Y range
Maximize mean of ranking according to SRD and ranking according to R^2	1.4621	0.4390	AROM area, AROM X range, AROM Y range, PROM area, PROM X range, PROM Y range, SPARC, mean velocity
Maximize R^2 subject to the constraint SRD ≤ 2	1.4621	0.4390	AROM area, AROM X range, AROM Y range, PROM area, PROM X range, PROM Y range, SPARC, mean velocity
Minimize SRD subject to $R^2 > 0.45$	2.6671	0.4551	AROM area, AROM X range, AROM Y range, PROM Area PROM X range, PROM Y range SPARC, mean velocity
Minimize SRD subject to $R^2 > 0.4$	1.4621	0.4390	AROM area, AROM X range, AROM Y range, PROM area, PROM X range, PROM Y range, SPARC, mean velocity
ARAT			
Maximize R^2	8.7030	0.4987	AROM AREA, AROM X range, AROM Y range
Maximize mean of ranking according to SRD and ranking according to R^2	2.7249	0.4410	AROM area, AROM X range, AROM Y range
Maximize R^2 subject to the constraint SRD ≤ 2	1.8067	0.3088	AROM area, AROM X range, AROM Y range, PROM Area, PROM X range
Minimize SRD subject to $R^2 > 0.45$	4.0953	0.4867	AROM area, AROM X range
Minimize SRD subject to $R^2 > 0.4$	2.6803	0.4246	AROM area, AROM X range, AROM Y range

FMA-UE: Fugl-Meyer Assessment – Upper Extremity; ARAT: Action Research Arm Test.

with or better than what could be expected based on previous work with similar methods. For example, Bosecker et al.[7] used multilinear regression models to estimate FMA scores in individuals with chronic stroke based on robotic measures and obtained a R value of 0.427 (i.e. R^2 of 0.18).

The lack of a relationship between SRD and predictive performance (R^2) is more surprising, but can be partly explained by the different data points used to compute the two metrics (see Figure 1(b)). In other words, models that were able to capture long-term trends in function (as evidenced by a high R^2) were not necessarily the same models that could consistently reproduce estimates over the seven closely spaced baseline assessments (as evidenced by a low SRD). Nonetheless, it was possible to find models that could satisfy both requirements to a reasonable degree. It does appear that the relationship between SRD and R^2 is partly dependent on the specific outcome measure, since a stronger relationship was found for the ARAT compared to the FMA-UE. This relationship is expected to depend on the outcome measure being predicted as well as the choice of robotic metrics, the modeling techniques used, and the design of the robotic device itself.

While several investigations of the reliability of upper limb robotic assessments have been conducted previously, their implications for guiding therapy progression have not been examined. Colombo et al.[29] quantified the intra-session and inter-session reliability of multiple individual robotic measures, finding high ICCs and minimal detectable differences (i.e. SRDs) that were smaller than the observed change over the course of a multi-week intervention. Keller et al.[30] measured inter- and intra-rater reliability of robotic measures using an upper-limb exoskeleton-type robot and a group of individuals with spinal cord injury. Although these results cannot be directly compared with our results from a population of stroke survivors using an end-effector robot, they found reliability values that ranged from poor to good depending on the metric. Their reliability analysis was based on ICC and correlation coefficients, which, as demonstrated here, are not necessarily indicative of SRD values. Several studies have described moderate to excellent inter-rater reliability depending on the measure.[31–33] None of these studies looked at the reliability of clinical outcome measure scores predicted from the robotic assessments, or considered the application of the measures for directing day-to-day therapy progression.

One limitation of this study is the sample size available, which was limited to nine individuals. Considering the nine planned assessments per participant and the use of leave-one-subject-out cross-validation, and accounting for missing data points, each model was trained on approximately 68 data points on average. It is possible that a larger dataset would have impacted on the robustness and properties of the models obtained. On the other hand, the availability of seven baseline assessments per participant from which to compute the SRDs and ICCs is a strength of the study, since few clinical studies are designed to include so many measurements during a stable period. The fact that the participants had chronic stroke is beneficial for the purposes of this study, because the multiple repeated baseline assessments in the pre-intervention phase could be conducted without the confounding influence of on-going recovery. The SRDs obtained would be expected to remain valid if the same measures were applied to individuals in the sub-acute stage.

Another limitation related to the sample size is the modeling technique chosen. Here we used multiple linear regression. Recent studies have investigated the use of nonlinear models to predict outcome measures from robotic assessments,[10,15] and demonstrated improved performance. However, as the complexity of the model increases, so does the risk of overfitting, and a larger dataset becomes necessary to ensure that the models obtained can generalize well. In this study, we judged that the amount of data available was not sufficient to support the development of nonlinear models. Doing so on a larger dataset would result in models with different properties. Nonetheless, our conclusions are based on a large number of linear models (16,383). While nonlinear models may result in higher R^2 values, we do not believe that there is any reason to expect that they would introduce stronger relationships between the model output SRDs and the model R^2s or input variable SRDs. It is important to keep in mind that the SRDs are not solely a result of the mathematical model, but are also influenced by the choice of metrics and the data collection procedures. For example, ROM in an end-effector robot may be difficult to reproduce very exactly day after day, because it may be partly influenced by small changes in the participant's positioning with respect to the table. These types of considerations will be reflected in the SRDs regardless of the modeling technique used.

Conclusion

If upper limb rehabilitation robots can estimate the user's function with a measurement error that is lower than the expected clinical changes on a timescale of days, then these robotic measures could be used to inform treatment decisions. This study has demonstrated that predictive models can be constructed that have suitable SRDs for this application. However, models with the highest predictive performance are not always the ones with the smallest SRDs, and vice

versa. We recommend that future work on using robotic rehabilitation devices for assessment explicitly include SRD evaluations.

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: Wang reports grants from Drummond Foundation, outside the submitted work. The robots used in this study have been provided in kind from an industry partner, Quanser Inc. The industry partner is not involved in the research studies, does not have access to study data, and there are no restrictions imposed on reporting of the data.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors gratefully acknowledge the funding from the Drummond Foundation (2014RFA#5), as well as the following individuals for their contributions to the data collection: Debbie Hebert, Cynthia Ho, Bing Ye, Scott Tomlinson and Aishwarya Sudhama.

Guarantor

JZ.

Contributorship

JZ and RW conceived the study. MC and RHW were involved in protocol development, gaining ethical approval, and data collection. JZ, MM and MC conducted the data analysis. JZ wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Acknowledgements

We would like to thank all of the individuals who participated in the study.

ORCID iD

José Zariffa http://orcid.org/0000-0002-8842-745X.

References

1. Krebs HI and Volpe BT. Rehabilitation robotics. *Handbook Clin Neurol* 2013; 110: 283–294.
2. Veerbeek JM, et al. Effects of robot-assisted therapy for the upper limb after stroke a systematic review and meta-analysis. *Neurorehabil Neural Repair* 2017; 31: 107–121.
3. Balasubramanian S, et al. Robotic assessment of upper limb motor function after stroke. *Am J Phys Med Rehabil* 2012; 91(11 Suppl 3): 255.
4. Rohrer B, et al. Movement smoothness changes during stroke recovery. *J Neurosci* 2002; 22: 8297–8304.
5. Chang JJ, et al. The constructs of kinematic measures for reaching performance in stroke patients. *J Med Biol Eng* 2008; 28: 65–70.
6. Colombo R, et al. Assessing mechanisms of recovery during robot-aided neurorehabilitation of the upper limb. *Neurorehabil Neural Repair* 2008; 22: 50–63.
7. Bosecker C, et al. Kinematic robot-based evaluation scales and clinical counterparts to measure upper limb motor performance in patients with chronic stroke. *Neurorehabil Neural Repair* 2010; 24: 62–69.
8. Celik O, et al. Normalized movement quality measures for therapeutic robots strongly correlate with clinical motor impairment measures. *IEEE Trans Neural Syst Rehabil Eng* 2010; 18: 433–444.
9. Gilliaux M, et al. Using the robotic device REAplan as a valid, reliable, and sensitive tool to quantify upper limb impairments in stroke patients. *J Rehabil Med* 2014; 46: 117–125.
10. Mostafavi SM, et al. Robot-based assessment of motor and proprioceptive function identifies biomarkers for prediction of functional independence measures. *J Neuroeng Rehab* 2015; 12: 1.
11. Semrau JA, et al. Examining differences in patterns of sensory and motor recovery after stroke with robotics. *Stroke* 2015; 46: 3459–3469.
12. Krebs HI, et al. Robot-aided neurorehabilitation: from evidence-based to science-based rehabilitation. *Top Stroke Rehabil* 2002; 8: 54–70.
13. McKenzie A, et al. Validity of robot-based assessments of upper extremity function. *Arch Phys Med Rehabil* 2017; 98: 1969–1976. e2.
14. Colombo R, et al. Modeling upper limb clinical scales by robot-measured performance parameters. In: *IEEE international conference on rehabilitation robotics (ICORR)*, Zurich, Switzerland, 29 June–1 July 2011.
15. Krebs HI, et al. Robotic measurement of arm movements after stroke establishes biomarkers of motor recovery. *Stroke* 2014; 45: 200–204.
16. Semrau JA, et al. Relationship between visuospatial neglect and kinesthetic deficits after stroke. *Neurorehabil Neural Repair* 2015; 29: 318–328.
17. Cortes JC, et al. A short and distinct time window for recovery of arm motor control early after stroke revealed with a global measure of trajectory kinematics. *Neurorehabil Neural Repair* 2017; 31: 552–560.
18. Massie CL, et al. A clinically relevant method of analyzing continuous change in robotic upper extremity chronic stroke rehabilitation. *Neurorehabil Neural Repair* 2016; 30(8): 703–712.
19. Beckerman H, et al. Smallest real difference, a link between reproducibility and responsiveness. *Qual Life Res* 2001; 10: 571–578.
20. Gowland C, et al. Measuring physical impairment and disability with the Chedoke-McMaster Stroke Assessment. *Stroke* 1993; 24: 58–63.
21. Lu EC, et al. Development of a robotic device for upper limb stroke rehabilitation: a user-centered design approach. *Paladyn* 2011; 2: 176–184.
22. Huq R, et al. Development of a portable robot and graphical user interface for haptic rehabilitation exercise. In: *4th IEEE RAS & EMBS international conference on*

biomedical robotics and biomechatronics (BioRob), Rome, Italy, 24–27 June 2012.
23. Huq R, et al. Development of a fuzzy logic based intelligent system for autonomous guidance of post-stroke rehabilitation exercise. In: *IEEE international conference on rehabilitation robotics (ICORR)*, Seattle, WA, USA, 24–26 June 2013.
24. Fugl-Meyer AR, et al. The post-stroke hemiplegic patient. 1. A method for evaluation of physical performance. *Scand J Rehabil Med* 1975; 7: 13–31.
25. Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. *Int J Rehab Res* 1981; 4: 483–492.
26. Zariffa J, et al. Relationship between clinical assessments of function and measurements from an upper-limb robotic rehabilitation device in cervical spinal cord injury. *IEEE Trans Neural Syst Rehabil Eng* 2012; 20: 341–350.
27. Balasubramanian S, et al. On the analysis of movement smoothness. *J Neuroeng Rehab* 2015; 12: 112.
28. Shrout PE and Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979; 86: 420.
29. Colombo R, et al. Test–retest reliability of robotic assessment measures for the evaluation of upper limb recovery. *IEEE Transac Neural Syst Rehab Eng* 2014; 22: 1020–1029.
30. Keller U, et al. Robot-assisted arm assessments in spinal cord injured patients: a consideration of concept study. *PLoS One* 2015; 10: e0126948.
31. Coderre AM, et al. Assessment of upper-limb sensorimotor function of subacute stroke patients using visually guided reaching, *Neurorehabil. Neural Repair* 2010; 24(6): 528–541.
32. Dukelow SP, et al. Quantitative assessment of limb position sense following stroke, *Neurorehabil. Neural Repair* 2010; 24(2): 178–187.
33. Semrau JA, et al. Inter-rater reliability of kinesthetic measurements with the KINARM robotic exoskeleton, *Journal of Neuroengineering and Rehabilitation* 2017; 14(1): 42.