ASNR AMERICAN SOCIETY OF NEUROREHABILITATION

Neurorehabilitation and
Neural Repair
2020, Vol. 34(12) 1078–1087

DOI: 10.1177/1545968320962483
journals.sagepub.com/home/nnr
SAGE

Improving Accelerometry-Based Measurement of Functional Use of the Upper Extremity After Stroke: Machine Learning Versus Counts Threshold Method

Peter S. Lum, PhD[1,2], Liqi Shu, MD[3], Elaine M. Bochniewicz, PhD[1], Tan Tran, MS[1], Lin-Ching Chang, DSc[1], Jessica Barth, MS, OTR/L[2], and Alexander W. Dromerick, MD[2,4]

Abstract

Background. Wrist-worn accelerometry provides objective monitoring of upper-extremity functional use, such as reaching tasks, but also detects nonfunctional movements, leading to ambiguity in monitoring results. *Objective.* Compare machine learning algorithms with standard methods (counts ratio) to improve accuracy in detecting functional activity. *Methods.* Healthy controls and individuals with stroke performed unstructured tasks in a simulated community environment (Test duration = 26 ± 8 minutes) while accelerometry and video were synchronously recorded. Human annotators scored each frame of the video as being functional or nonfunctional activity, providing ground truth. Several machine learning algorithms were developed to separate functional from nonfunctional activity in the accelerometer data. We also calculated the counts ratio, which uses a thresholding scheme to calculate the duration of activity in the paretic limb normalized by the less-affected limb. *Results.* The counts ratio was not significantly correlated with ground truth and had large errors ($r = 0.48$; $P = .16$; average error = 52.7%) because of high levels of nonfunctional movement in the paretic limb. Counts did not increase with increased functional movement. The best-performing intrasubject machine learning algorithm had an accuracy of 92.6% in the paretic limb of stroke patients, and the correlation with ground truth was $r = 0.99$ ($P < .001$; average error = 3.9%). The best intersubject model had an accuracy of 74.2% and a correlation of $r = 0.81$ ($P = .005$; average error = 5.2%) with ground truth. *Conclusions.* In our sample, the counts ratio did not accurately reflect functional activity. Machine learning algorithms were more accurate, and future work should focus on the development of a clinical tool.

Keywords

upper extremity, stroke, neurorehabilitation, accelerometry, machine learning

Introduction

More than 795 000 individuals have a stroke each year in the United States.[1] Based on data from multiple cohorts,[2-6] up to 77% will have persistent upper-extremity (UE) impairment. Yet the development of markedly effective treatments has been slow[7,8] and uncertain.[9,10] An important reason for the lack of efficacy findings in clinical trials is the lack of a direct measure of the outcome of interest: productive functional use of the UE in everyday life. Without such a measure, clinicians and clinical trialists simply cannot know for certain if their treatments worked. Current approaches settle for use of proxies: self-report of UE use[11,12] (with its attendant biases) and motor performance measures in the laboratory and clinic[13,14] (with uncertain correspondence to everyday life). Thus, there is a clinical and research need for a quantitative, objective, and psychometrically sound measure of UE productive use in the community.

There is now a large body of work on wearable sensors that can track UE activity, with the majority of studies using wrist-worn accelerometers and the "counts threshold method" for detecting movement.[15-18] This method estimates the total amount of time the UE is in motion. Several studies have reported that the counts threshold method often correlates significantly with other clinical

[1]The Catholic University of America, Washington, DC, USA
[2]MedStar National Rehabilitation Network, Washington, DC, USA
[3]Warren Alpert Medical School of Brown University, Providence, RI, USA
[4]Georgetown University School of Medicine, Washington, DC, USA

Corresponding Author:
Peter S. Lum, PhD, Biomedical Engineering, The Catholic University of America, Pangborn Hall, Room 131, 620 Michigan Ave NE, Washington, DC 20064, USA.
Email: lum@cua.edu

Table 1. Stroke Participant Demographic Data.

Participant no.	Age (years)	Sex	Affected limb	Stroke	Location	Months poststroke	ARAT
1	77	M	Right	Embolic	Cerebrum	23	41
2	35	M	Left	Embolic	Not available	35	23
3	56	M	Left	Ischemic	Basal ganglia	17	19
4	49	F	Left	Ischemic	Basal ganglia	19	20
5	57	M	Right	Hemorrhagic	Basal ganglia	104	16
6	63	M	Right	Ischemic	Temporal lobe, thalamus	77	32
7	47	F	Right	Ischemic	Pontine	1	33
8	50	M	Right	Ischemic	Not available	53	15
9	66	M	Right	Ischemic	Corona radiata	69	5
10	65	M	Right	Hemorrhagic	Frontal and occipital lobes	20	31

Abbreviations: ARAT, Action Research Arm Test; F, female; M, male.

scales.[19-22] However, the correlations reported are often weak. For example, the largest sample was collected as part of the EXCITE clinical trial (n = 169), and the reported *r* value was only 0.52 between accelerometry and the Motor Activity Log (MAL).[23] Responsiveness to change is mixed: some studies report changes in accelerometry that parallel change in clinical scores,[24-29] whereas other studies have reported no changes in accelerometer-based metrics even as clinical scores improved.[30-33]

There are significant differences in what is measured by self-report scales of functional use (MAL) and acwcelerometry.[34] Accelerometry measures both functional activity (ie, reaching to grasp, gesturing, balancing functions) and nonfunctional movements associated with gait and whole body movements. A bus ride will lead to many instances of acceleration not associated with functional movement. The effects of nonfunctional movements are thought to be eliminated by normalizing duration of movement in the paretic limb by duration in the less-affected limb, referred to as the *counts ratio*. However, there have been no prior studies to confirm that this normalization is effective. In fact, there have been surprisingly few attempts to quantify if any of the accelerometry-based metrics actually measure functional use accurately. Early attempts at validation focused only on confirming that the duration of movement via the thresholding method correlates with video annotation[35] or examined the correlation between duration of movement within 15-minute blocks and a video-based metric that combined the amount of functional and nonfunctional movements into a single average score.[36]

We have developed machine learning algorithms for directly measuring amount of functional movement using accelerometry. In previous work, we determined the accuracy of a single machine learning algorithm during a session of unstructured activities using ground truth from a video-based method that provides frame-by-frame scoring of the activities.[37] In this study, using the same data set, we investigated several alternative machine learning algorithms to find generalizable models for our application and have improved accuracies compared with previous reports. We have extended the analysis of this data set by directly comparing the counts threshold method and our machine learning algorithms against ground truth from video annotation. Furthermore, we also analyzed data collected from the dominant limb of controls and the less-affected limb of stroke participants (not analyzed previously), so that ratio metrics could be studied. Most important for neurorehabilitation, we present the first study to separate activity measured by the counts threshold method into functional and nonfunctional categories and present machine learning algorithms to automatically identify periods of functional movement from accelerometer data.

Methods

Participants

Data collection procedures were described previously.[37] A total of 10 healthy controls (4 male, 6 female; 43 ± 15.9 years old) and 10 individuals with stroke (8 male, 2 female; 56 ± 10.4 years old) participated (Table 1). The controls had no self-reported injuries that would alter or impair their use of either UE. Inclusion criteria for the stroke participants were the following: (1) ischemic or hemorrhagic stroke; (2) Mini-Mental Status Examination score >24[38]; and (3) no UE conditions that limited use prior to the stroke. Participants were excluded if they exhibited neglect during a clinical examination. All controls and stroke patients (prior to stroke) were right-hand dominant (Edinburgh Inventory[39]).

Measures

A custom-designed wrist worn inertial measurement unit (IMU) was developed that collected linear acceleration in 3 axes at 200 Hz.[40] The standard counts method we adopted only uses linear acceleration, so the machine learning was applied only to the acceleration data from the IMU. This guaranteed that any advantage of the machine learning was not a result of a larger input data set that included data from the gyroscope and magnetometer on the IMU. This also allows application of the machine learning to other accelerometry data sets that do not include the gyroscope and magnetometer data. Sensors were placed on both wrists. The machine learning algorithms were based only on 1 IMU, but data was collected on both wrists to allow comparison between limbs and to calculate ratio scores. The Action Research Arm Test (ARAT)[41] was used to assess the functional limitations of the paretic limb of stroke participants (mean = 23.5 ± 10.7).

Procedures

To obtain a realistic environment for data collection, we used the Independence Square facility at MedStar National Rehabilitation Hospital. The facility includes a completely functional kitchen, bedroom, a store for shopping activities, and a car to practice transfers. Participants were instructed to perform 4 typical instrumental activities of daily living. (1) In the laundry activity, participants moved clothes from a closet, placed them in a washer, moved the clothes to the dryer, and folded or hung the clothes on hooks in the closet. (2) In the kitchen activity, participants loaded and unloaded the dishwasher, cut an apple, picked up items on the floor, and used a broom to sweep the floor. (3) In the shopping activity, participants transferred into and out of the car, gathered grocery items from the store and placed them into the car, then removed them from the car. (4) In the bed making activity, participants removed the sheets and pillowcases from a bed and then replaced them. Participants were instructed to perform activities as they would naturally do at home or in the community. No specific instructions were given as to which arm to use for any task. The only exceptions were that participants were instructed to gather groceries with 1 hand and move a large box with 2 hands. Between these 4 activities, participants sat and experimenters engaged them in conversation or walked around the facility (approximately 10 minutes of walking). There was no set time limit to complete the activities. Participants wore the sensors throughout the experiment, which was also videotaped (30 Hz). The device weighed less than 0.5 lb, and none of the participants noted that the device interfered with the activities. An occupational therapist observed all of the data collection and reported that the device did not interfere with activities.

Data Analysis

The video was annotated using a method described previously.[37] Briefly, 3 annotators unrelated to the study were trained on the Functional Arm Activity Behavioral Observation System (FAABOS).[36] Annotators watched the video in real time, and when an arm movement was seen, the video was stopped and rewound to mark the start and end of the movement. All frames between the start and end frame of each movement were labeled according to the 5 FAABOS categories. We subsequently collapsed these into 3 categories: functional, nonfunctional, or unknown by each annotator. The functional category included gesturing, reach to grasp, pushing open a door, and so on. The nonfunctional category included arm movements associated with gait, sit-to-stand, or other whole body movement that did not include a functional arm movement. The no-movement periods between these periods of arm movement were also labeled nonfunctional. Arm swing is important for balance during gait, but we elected to delegate these movements as nonfunctional in order to create a contrast with volitional prehensile movements, which are the traditional targets of UE rehabilitation. Additionally, detection of arm swing during gait is a common criticism of the counts thresholding method, with some researchers adding a sensor on the leg to detect gait movements so that gait-related arm movements can be neglected.[31] A major goal was automatic separation of whole body and gait-related movements from reach and grasp movements. The final categorization was determined by majority vote. About 3% of the data were labeled as unknown because the arm was occluded in the video or there was no majority consensus across the annotators.

The accelerometry and video data streams were synchronized at the start of collection, and we performed spot checks to make sure synchrony was maintained throughout by checking rest periods, which are easy to visually spot in both streams. Each 4-s block of accelerometer data was given a ground truth label (functional, nonfunctional, mixed, or unknown) based on the corresponding video frame annotations. Blocks with a majority of unknown frames were excluded from analysis. Mixed blocks had a combination of functional and nonfunctional video frames and were labeled based on the majority of frames in the block. We then computed 11 features from each 4-s epoch of sensor data. The mean and variance across the epoch were computed for *x, y*, and *z* acceleration components. The Euclidean norm was calculated across the 3 dimensions at each sample point ($\sqrt{x^2+y^2+z^2}$). The mean, variance, minima, and maxima of the Euclidean norm across the 4-s period were also computed. The Shannon entropy was calculated as $-\sum_{i=1}^{N} p_i \log(p_i)$, where N is the number of sample points in the 4-s block and p_i is the probability of occurrence of each value in the block, based on a kernal density estimator applied to all values in the 4-s block with a Gaussian smoothing function. Higher entropy implies more uncertainty or unpredictability in the data. These features were chosen because of their simplicity and were proven in our prior study to give good results.

Feature normalization using min-max scaling was performed before classification. We investigated the unsupervised K-means clustering algorithm and 4 supervised algorithms: K-Nearest Neighbors (KNN), Random Forest, Linear Support Vector Machine (SVM), and Radial Basis Function SVM (RBF SVM). Because of the heterogeneity of stroke survivors, we believed that the ensemble method (Random Forest) would perform well. SVM is considered one of the most robust and accurate classification algorithms, whereas KNN tends to perform well when the number of data points per subject is large and the number of features is small. We also tested if performing principal component analysis (PCA) on our feature data set affected the classification results. Given an expectedly high variability between limbs in stroke patients and between stroke patients and controls, different models were built for each limb: dominant control, nondominant control, paretic stroke, and less-affected stroke. Both intersubject and intrasubject models were built for each limb. For the intrasubject models, stratified 5-fold cross-validation was used. For the intersubject models, leave-one-out cross-validation (ie, 10-fold in our case) was used because it is approximately unbiased and is the best choice when the number of subjects is small.[42] Note that the mixed blocks were excluded during the training to improve the model performance but included during the testing phase to reflect the real-world situation. For each machine learning algorithm, there are model hyperparameters, such as C, kernel, and gamma for SVM.[43] Exhaustive grid search was conducted to determine the optimal set of hyperparameters for each algorithm (Python programing language with Scikit-learn, Pandas, and Matplotlib libraries).[43] For example, the RBF SVM grid search parameters were C [1, 10, 100, 1000] and gamma [0.001, 0.0001]. Parameters are defined in the user manual of Scikit-learn.

For each algorithm, the classification accuracy was defined as the percentage of the data correctly classified into functional or nonfunctional categories using the human annotation as the ground truth. For each limb, we also calculated the %functional, defined as the total duration of time in functional movement, normalized by the total duration of the trial. This metric was calculated directly from video annotations and using both the intersubject and intrasubject algorithms. For these calculations, the mixed blocks were included.

Counts Threshold Method

As a comparison method, we processed the data via the counts threshold method first proposed by Uswatte et al[35] and later detailed by Urbin et al.[27] The ActiGraph family of wrist worn sensors are used, which samples acceleration at 30 Hz. An offline proprietary filter reduces the data into "counts" over each 1-s epoch. The filter removes the effects of gravity and higher frequencies not present during human movement.[44] The counts unit is proportional to the acceleration magnitude within that period. The data sampling rate was reduced from 200 to 30 Hz via interpolation and formatted so that the ActiLife software would accept the data and calculate the counts metric. The 3 axes were combined via the Euclidean norm into a single counts value for each 1-s epoch.[27] Metrics calculated included the following: (1) usage: sum of 1-s epochs where the counts were >1, normalized by the total number of epochs in the trial; (2) usage ratio: usage in the paretic arm normalized by usage in the less-affected arm. ActiGraph is the accelerometer most commonly used in research studies,[45] and code is now available to calculate ActiGraph counts from raw accelerometer data from any sensor.[46]

To better understand the differences between functional and nonfunctional movements, the counts threshold method was used to mark 1-s epochs as active (counts > 1) or nonactive. After neglecting the nonactive epochs, we calculated the duration of functional and nonfunctional movement using the frame-by-frame annotation data. The movement count value from each active 1-s epoch was assigned to all video frames within that epoch. This then allowed calculation of the average acceleration amplitude (in counts) for functional and nonfunctional movement. Repeated-measures ANOVA was used to analyze these data. The within-subject factors were limb (dominant, nondominant) and movement type (functional, nonfunctional). Separate ANOVAs were performed for the stroke group and the control group. Significant effects were examined with *t*-tests with Bonferroni correction.

Results

Table 2 shows the classification accuracy of the machine learning algorithms. For the nondominant limb of controls, best accuracies were 96.0% and 91.1% for intrasubject and intersubject modeling, respectively. For the paretic limb of stroke patients, best accuracies were 92.6% and 74.2% for intrasubject and intersubject modeling, respectively. Random Forest performed better in intrasubject modeling, and RBF SVM performed better in the nondominant or paretic limb (intersubject modeling). Overall, the results from Random Forest and RBF SVM were comparable. PCA did not improve the results. To calculate %functional use for each limb, we selected only the Random Forest algorithm.

Figure 1A compares the results of the counts threshold method and the Random Forest machine learning algorithm. The ground truth %functional use values are available from the video annotation. When considering only the paretic limb of stroke participants, the intrasubject model estimates of %functional were highly correlated with ground truth video; the correlation coefficient was r =0.99

Table 2. Classification Accuracy.

	Intrasubject (percentage ± SD)		Intersubject (percentage ± SD)	
Algorithms	Control	Stroke	Control	Stroke
Nondominant or paretic limb				
K-Nearest Neighbors	95.17 ± 1.05	89.32 ± 7.53	90.45 ± 3.14	65.90 ± 8.54
Random Forest	96.05 ± 1.22	92.61 ± 3.51	88.27 ± 4.35	68.35 ± 8.08
Linear SVM	92.28 ± 1.95	85.52 ± 9.16	88.61 ± 3.59	70.41 ± 13.92
RBF SVM	94.59 ± 1.36	89.23 ± 6.83	91.07 ± 3.63	74.24 ± 11.43
K-means clustering	73.94 ± 4.52	67.80 ± 8.66	72.63 ± 5.62	59.12 ± 17.83
Dominant or less-affected limb				
K-Nearest Neighbors	95.18 ± 1.47	92.80 ± 7.22	91.18 ± 3.25	84.10 ± 11.39
Random Forest	96.64 ± 1.00	94.64 ± 4.57	90.52 ± 4.87	83.32 ± 12.05
Linear SVM	92.97 ± 2.19	91.29 ± 7.57	89.86 ± 4.45	84.90 ± 10.18
RBF SVM	93.38 ± 1.81	92.45 ± 7.35	90.83 ± 4.72	84.76 ± 12.02
K-means clustering	76.78 ± 4.02	83.20 ± 10.98	75.80 ± 4.68	83.05 ± 10.77

Abbreviations: RBF SVM, Radial Basis Function Support Vector Machine.

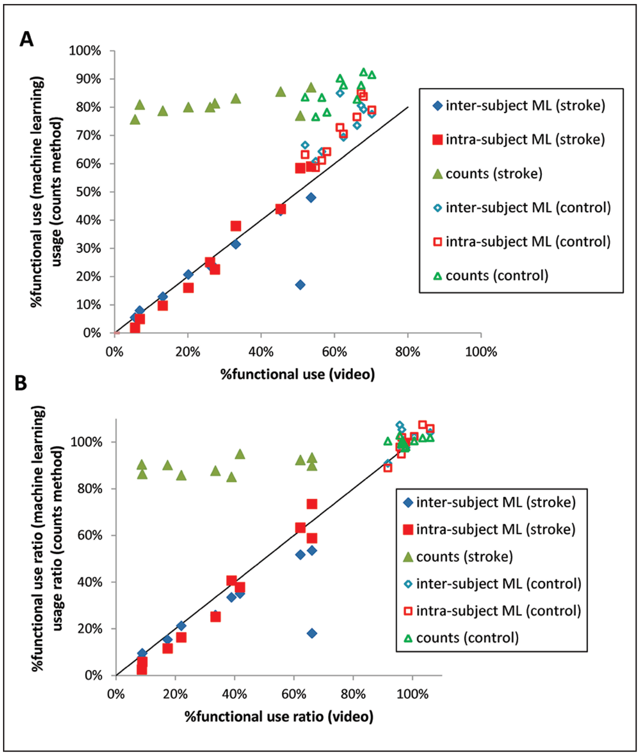

Figure 1. A. The %functional use from machine learning (ML) and usage from the counts threshold method compared with %functional use from video annotation. The paretic and nondominant limbs of controls are represented. The solid line is a reference line representing perfect correlation with ground truth. B. The %functional use ratio and usage ratio compared with %functional use ratio from video. Paretic limb values were normalized by values from the less-affected limb. In healthy controls, the nondominant limb was normalized by the dominant limb. The counts method used a 1-s epoch, and the machine learning was Random Forest using a 4-s epoch.

($P < .001$), and the error magnitude was 3.9% (Table 3). For intersubject modeling, the correlation dropped to r=0.81 ($P = .005$) and the error increased to 5.2%. Figure 1A also shows that usage calculated from the counts threshold method is a very poor representation of %functional; despite large variability across participants in %functional use as determined by video ground truth, the usage metric measured almost no differences across participants. The correlation of usage with %functional (video) was not significant ($r = 0.57$; $P = .085$), and the error magnitude was 52.7%.

Calculating the ratio of usage between the paretic and less-affected limb ("usage ratio") is a common approach to reduce the effects of nonfunctional movements, such as arm swing during gait and whole body movements. The results did not change significantly when metrics were normalized by metrics from the less-affected limb (Figure 1B). The intrasubject model again had the highest correlation with ground truth and lowest error, and the usage ratio continued to grossly overestimate the amount of functional movement (see Table 3 for complete details).

One participant did very poorly in intersubject modeling of the paretic limb, suggesting that their movement patterns differed from those of the other stroke participants. Figure 2 shows the ground truth and predictions from the intersubject modeling of this participant using the Random Forest algorithm. From these data, we flagged instances of false positives (predicted functional movement when there was none) and false negatives (missed instances of functional movement). There were large incidences of false negatives, which contributed to the underestimation of functional use in this participant. Figure 2 also shows the results of the intrasubject modeling applied to this same participant using the same algorithm. The intrasubject modeling greatly reduced the false negatives in the prediction and improved the accuracy.

Table 3. Correlations and Errors of Metrics Versus Functional Use From Video (Stroke Data Only).

	Paretic limb			Ratio of paretic/less-affected limb		
Metrics	*r* Value	*P*	Error	*r* Value	*P*	Error
%functional (Intrasubject)	0.99	<.001	3.9%	0.99	<.001	5.1%
%functional (Intersubject)	0.81	.005	5.2%	0.78	.008	9.6%
Usage (1 count)[a]	0.57	.085	52.7%	0.48	.16	53.0%
Usage (20 counts)	0.46	.18	26.6%	0.37	.30	29.7%
Usage (40 counts)	0.27	.45	15.0%	0.34	.34	19.6%

[a] count refers to the threshold used to separate movement from rest.

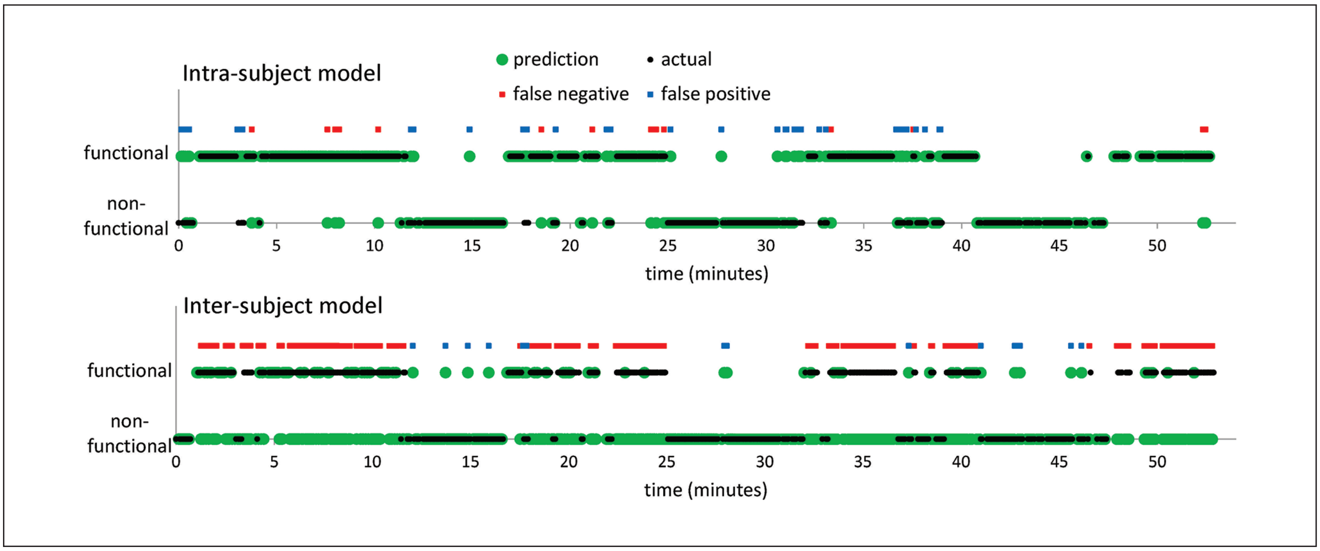

Figure 2. Intrasubject and intersubject model predictions (based on Random Forest) from the stroke participant with the worst intersubject model. Intersubject model errors are mostly false negatives (marked by red squares), which are reduced by the intrasubject modeling.

Because the usage metric appears to be detecting large amounts of nonfunctional movement, we explored the possibility that increasing the threshold in the counts threshold method might improve the result by rejecting slower movements, which might be biased toward nonfunctional movements. The error magnitude between usage and ground truth %functional movement does decrease with higher thresholds; however, the correlations with ground truth also decrease because of increased scatter in the data (see Table 3 for complete details). All correlations remained nonsignificant after increasing the counts threshold.

Combining the video annotation and the counts data allows elimination of periods where there was no movement (counts < 1). Figure 3A shows the percentage of time in functional and nonfunctional movement (after eliminating no movement periods) in both limbs of stroke and control participants. For stroke participants, the repeated-measures ANOVA reported a significant interaction between limb and movement type ($P < .001$). For controls, the ANOVA reported a significant effect of movement type ($P < .001$). A series of within-subject *t*-tests (with Bonferroni correction) found the following differences. The %nonfunctional movement was significantly lower than %functional movement in the less-affected limb of stroke participants ($P = .001$) and in both limbs of controls ($P < .001$). In contrast, the %nonfunctional movement was significantly higher than %functional movement in the paretic limb ($P = .007$). When comparing across limbs in the same participant, the %nonfunctional movement was significantly higher in the paretic limb compared with the less-affected limb ($P < .001$), whereas %functional movement was significantly lower in the paretic limb compared with the less-affected limb ($P < .001$). There were no differences between limbs in controls ($P > .99$).

We also calculated the average amplitude (counts) when the arm was in motion. Nonfunctional movements were slower than functional arm movements in the less-affected arm of stroke participants and both arms of

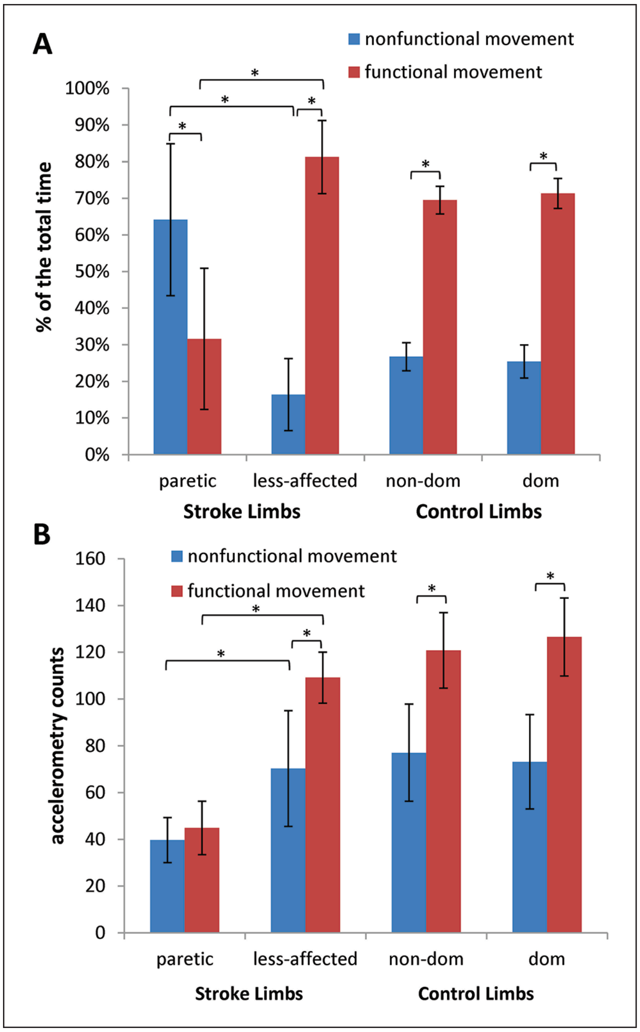

Figure 3. (A) Percentage of time spent and (B) average of accelerometry counts in nonfunctional and functional movement in stroke participants and controls. Periods of no movement were removed before this analysis.
Abbreviation: dom, dominant.
*Significant difference, $P < .05$.

controls ($P < .001$; Figure 3B). This was not true for the paretic limb, where there were no differences in acceleration amplitude between functional and nonfunctional movements ($P > .99$). This explains the inability of the thresholding method to separate functional and nonfunctional movements in the paretic limb, even as the threshold was varied. When comparing the 2 limbs from the same participants, accelerations were lower in the paretic limb compared with the less-affected limb for both functional ($P < .001$) and nonfunctional ($P = .007$) movements. In contrast, there were no differences in acceleration between limbs in the controls ($P > .99$).

Using the stroke data only, we calculated the correlations between ARAT scores and the ratio metrics. Using the video annotations, there was a significant correlation between the %functional use ratio and ARAT scores ($r = 0.72$; $P = .019$). Using the machine learning intrasubject modeling, there was a significant correlation between %functional use ratio and ARAT scores ($r = 0.79$; $P = .007$). However, the usage ratio calculated from the counts threshold method was not correlated with ARAT scores ($r = 0.36$; $P = .31$).

The machine learning results were not sensitive to our parameter selections. The Random Forest analysis was repeated for epoch lengths of 5, 4, 3, 2, and 1 s. In the intrasubject analysis, the highest accuracy was with a 4-s epoch and the lowest accuracy, with the 1-s epoch, but the difference was only 0.7% ± 1.8% ($P = .85$). Intersubject analysis also found no difference in accuracy between 4- and 1-s time epochs (Difference = 0.6% ± 1.3%; $P = .89$). After reducing the sample rate to 30 Hz, the accuracy of the Random Forest model increased by only 0.1% ± 0.9% ($P = .98$).

Discussion

We compared machine learning for extracting the amount of functional UE movement to the counts threshold method of detecting the amount of movement (usage). An important and novel result from this study is that the counts threshold method detects more nonfunctional movement than functional movement in the paretic limbs of stroke participants. The reverse was true for the less-affected limb and for both limbs of controls, where a much larger percentage of the movement detected was functional. As a result, the duration of movement from the counts threshold method is grossly higher than the amount of functional movement determined from video annotation. Normalizing by usage in the less-affected limb did not change these results, which refute the assumption that increased movement recorded via the counts threshold method is an objective measure of increased functional limb use. If present in a larger validation cohort, this result has importance in studies of UE recovery and suggests that the counts threshold method leads to substantial errors. We also further developed machine learning algorithms that extract functional movement patterns from the accelerometry data. K-means clustering performed poorly compared with the supervised algorithms. The assumptions that all features have equal importance for every cluster and that clusters are spherical and evenly sized might not be appropriate for this application. The supervised intrasubject modeling produced good results with estimates of %functional use that correlated strongly with ground truth. If usability can be improved, these machine learning algorithms can be an objective tool for measuring real-world UE use.

Whereas most accelerometry studies have focused on movement duration, acceleration magnitude could have interesting features. The density plot method developed by Bailey et al[47] showed that in both healthy controls and individuals with stroke, the most common movements were bilateral with fairly low acceleration magnitudes. Simultaneous UE activity made up 67% (7.2/10.7 hours) of total UE activity in nondisabled adults and 49% (4.1/8.4 hours) in individuals with stroke. Asymmetries in density plots and lower magnitudes were characteristic of individuals with stroke, indicating a shift toward less-affected arm use and slower movement speeds. These methods provide a rich data set that can be used to characterize the real-world behavior of individuals. The addition of machine learning algorithms could extend these methods by eliminating movement not associated with functional limb use.

The counts threshold method was pioneered by Uswatte et al.[35] In their seminal article, they reported a 98% agreement when 2-s epochs were tested for movement duration by the counts threshold method and by human observers of synchronous video in stroke patients performing 15 minutes of Activities of Daily Living (ADL).[35] However, the authors noted that duration of movement may not reflect duration of functional movement, which is more clinically relevant. Subsequent work from this group reported high test-retest reliability (r = 0.86) and an increase of 8% in the usage ratio after Constraint-Induced (CI) therapy.[28] This group also attempted to validate the usage ratio metric with video annotation using the FAABOS scoring method.[36] Participants performed activities in the home or clinic without any special instruction while being videotaped and wearing accelerometers. The correlation across participants between mean FAABOS rating over randomly selected 15-minute blocks and mean accelerometry ratio metric was r =0.55. Although this does provide some evidence of convergent validity between the usage ratio and a course assessment of amount of functional use (mean FAABOS score over the 15-minute period), there are important distinctions with our study that should be noted. We used the FAABOS scoring method to annotate each frame of the video as functional or nonfunctional, so that the total duration of functional movement could be determined from the video instead of relying on mean FAABOS score. This allows calculation of ground truth on %functional use for comparison to the accelerometry. Our machine learning algorithms separate functional and nonfunctional arm movement from the sensor data directly without the need for human observation. We are not aware of any comparable methods that have demonstrated this capability.

Several studies have reported relatively poor or nonexistent correlations between accelerometry metrics and clinical scales of function.[30-33] Function measured in the clinic is only one of many factors that can affect how and when the paretic limb is used in real-world settings. Other studies have reported that accelerometry metrics correlate weakly or not at all with self-report scales of UE use.[23,34] This might be a result of self-report bias or differences between the domains of total movement and functional use. Our results suggest that another factor in these poor correlations is the large amount of nonfunctional movement detected by the counts threshold method. Note that the usage ratio metric based on the counts threshold method did not correlate with ARAT scores in our sample, but the %functional from video annotation and machine learning both correlated significantly with ARAT scores.

Limitations

The usage and functional use percentages reported here should not be extrapolated to 24-hour recording periods. Our activity script has a higher percentage of activities than what would be expected from a full day of recording where more periods of inactivity or rest would be expected. This might explain why the usage ratio we observed is >80% in all participants, whereas this ratio has been reported to be 56% in other studies.[23] Additionally, it is unknown how the machine learning will respond to activities not in our activity script. Although we believe that our selected activities are a good representation of a fairly wide range of activities, the algorithm accuracy may decrease with untested activities. Tasks such as typing might be missed and better captured by devices that record finger movement directly.[48,49] The intrasubject models performed the best, but they represent a significant burden because our activity script is long (mean = 33 minutes), and a total of 6.9 person-hours were required to fully annotate the movements of each participant. We are researching if a reduced activity script is sufficient. The intersubject modeling had 1 clear outlier. This outlier had the most functional ability of the stroke patients, with an ARAT score of 41, whereas the remaining participants had ARAT scores of 33 or less. Qualitative examination of the video from this participant found that his shoulder and elbow movements seemed normal when compared with controls, likely leading to the poor intersubject models. We are currently investigating if intersubject model accuracy can be improved by incorporating the functional level of patients.

Conclusions

The error associated with the standard counts threshold analyses of accelerometric data obtained in stroke participants may be much larger than previously thought. We advise caution when using the counts threshold method, which detects large amounts of nonfunctional movement in the paretic limb during activities. Machine learning algorithms based on intrasubject modeling can accurately measure the amount of functional movement during unscripted ADL.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Department of Health and Human Services (NIDILRR; Grant Number 90REGE0004); the US Army Medical Research and Material Command (Grant Number W81XWH-11-1-0632); the Department of Veterans Affairs (Grant Number B7510R); and The MITRE Corporation.

ORCID iDs

Peter S. Lum https://orcid.org/0000-0002-4735-6114

Jessica Barth https://orcid.org/0000-0001-8051-9539

References

1. Benjamin EJ, Virani SS, Callaway CW, et al. Heart Disease and Stroke Statistics—2018 update: a report from the American Heart Association. *Circulation*. 2018;137:e67-e492.
2. Lai SM, Studenski S, Duncan PW, Perera S. Persisting consequences of stroke measured by the Stroke Impact Scale. *Stroke*. 2002;33:1840-1844.
3. Gresham GE, Phillips TF, Wolf PA, McNamara PM, Kannel WB, Dawber TR. Epidemiolgic profile of long-term stroke disability: the Framingham study. *Arch Phys Med Rehabil*. 1979;60:487-491.
4. Parker VM, Wade DT, Langton-Hewer R. Loss of arm function after stroke: measurement, frequency, and recovery. *Int Rehabil Med*. 1986;8:69-73.
5. Wade DT, Langton-Hewer R, Wood VA, Skilbeck CE, Ismail HM. The hemiplegic arm after stroke: measurement and recovery. *J Neurol Neurosurg Psychiatry*. 1983;46:521-524.
6. Lawrence ES, Coshall C, Dundas R, et al. Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. *Stroke*. 2001;32:1279-1284.
7. Langhorne P, Coupar F, Pollock A. Motor recovery after stroke: a systematic review. *Lancet Neurol*. 2009;8:741-754.
8. Veerbeek JM, van Wegen E, van Peppen R, et al. What is the evidence for physical therapy poststroke? A systematic review and meta-analysis. *PLoS One*. 2014;9:e87987.
9. AVERT Trial Collaboration Group. Efficacy and safety of very early mobilisation within 24 h of stroke onset (AVERT): a randomised controlled trial. *Lancet*. 2015;386:46-55.
10. Dromerick AW, Lang CE, Birkenmeier RL, et al. Very Early Constraint-Induced Movement during Stroke Rehabilitation (VECTORS): a single-center RCT. *Neurology*. 2009;73: 195-201.
11. Uswatte G, Taub E, Morris D, Light K, Thompson PA. The Motor Activity Log-28: assessing daily use of the hemiparetic arm after stroke. *Neurology*. 2006;67:1189-1194.
12. Duncan PW, Wallace D, Lai SM, Johnson D, Embretson S, Laster LJ. The Stroke Impact Scale Version 2.0: evaluation of reliability, validity, and sensitivity to change. *Stroke*. 1999;30:2131-2140.
13. Edwards DF, Lang CE, Wagner JM, Birkenmeier R, Dromerick AW. An evaluation of the Wolf Motor Function Test in motor trials early after stroke. *Arch Phys Med Rehabil*. 2012;93:660-668.
14. van der Lee JH, Beckerman H, Lankhorst GJ, Bouter LM. The responsiveness of the Action Research Arm test and the Fugl-Meyer Assessment scale in chronic stroke patients. *J Rehabil Med*. 2001;33:110-113.
15. Dobkin BH, Martinez C. Wearable sensors to monitor, enable feedback, and measure outcomes of activity and practice. *Curr Neurol Neurosci Rep*. 2018;18:87.
16. Gebruers N, Vanroy C, Truijen S, Engelborghs S, De Deyn PP. Monitoring of physical activity after stroke: a systematic review of accelerometry-based measures. *Arch Phys Med Rehabil*. 2010;91:288-297.
17. Lang CE, Wagner JM, Edwards DF, Dromerick AW. Upper extremity use in people with hemiparesis in the first few weeks after stroke. *J Neurol Phys Ther*. 2007;31:56-63.
18. Noorkoiv M, Rodgers H, Price CI. Accelerometer measurement of upper extremity movement after stroke: a systematic review of clinical studies. *J Neuroeng Rehabil*. 2014;11:144.
19. Gebruers N, Truijen S, Engelborghs S, Nagels G, Brouns R, De Deyn PP. Actigraphic measurement of motor deficits in acute ischemic stroke. *Cerebrovasc Dis*. 2008;26:533-540.
20. Shim S, Kim H, Jung J. Comparison of upper extremity motor recovery of stroke patients with actual physical activity in their daily lives measured with accelerometers. *J Phys Ther Sci*. 2014;26:1009-1011.
21. Thrane G, Emaus N, Askim T, Anke A. Arm use in patients with subacute stroke monitored by accelerometry: association with motor impairment and influence on self-dependence. *J Rehabil Med*. 2011;43:299-304.
22. van der Pas SC, Verbunt JA, Breukelaar DE, van Woerden R, Seelen HA. Assessment of arm activity using triaxial accelerometry in patients with a stroke. *Arch Phys Med Rehabil*. 2011;92:1437-1442.
23. Uswatte G, Giuliani C, Winstein C, Zeringue A, Hobbs L, Wolf SL. Validity of accelerometry for monitoring real-world arm activity in patients with subacute stroke: evidence from the extremity constraint-induced therapy evaluation trial. *Arch Phys Med Rehabil*. 2006;87:1340-1345.
24. Chen HL, Lin KC, Hsieh YW, Wu CY, Liing RJ, Chen CL. A study of predictive validity, responsiveness, and minimal clinically important difference of arm accelerometer in real-world activity of patients with chronic stroke. *Clin Rehabil*. 2018;32:75-83.
25. Liao WW, Wu CY, Hsieh YW, Lin KC, Chang WY. Effects of robot-assisted upper limb rehabilitation on daily function and real-world arm activity in patients with chronic stroke: a randomized controlled trial. *Clin Rehabil*. 2012;26:111-120.
26. Taub E, Uswatte G, Bowman MH, et al. Constraint-induced movement therapy combined with conventional neurorehabilitation techniques in chronic stroke patients with plegic hands: a case series. *Arch Phys Med Rehabil*. 2013;94:86-94.
27. Urbin MA, Waddell KJ, Lang CE. Acceleration metrics are responsive to change in upper extremity function of stroke survivors. *Arch Phys Med Rehabil*. 2015;96:854-861.

28. Uswatte G, Foo WL, Olmstead H, Lopez K, Holand A, Simms LB. Ambulatory monitoring of arm movement using accelerometry: an objective measure of upper-extremity rehabilitation in persons with chronic stroke. *Arch Phys Med Rehabil*. 2005;86:1498-1501.
29. Uswatte G, Taub E, Morris D, Vignolo M, McCulloch K. Reliability and validity of the upper-extremity Motor Activity Log-14 for measuring real-world arm use. *Stroke*. 2005;36:2493-2496.
30. Doman CA, Waddell KJ, Bailey RR, Moore JL, Lang CE. Changes in upper-extremity functional capacity and daily performance during outpatient occupational therapy for people with stroke. *Am J Occup Ther*. 2016;70:7003290040p1-7003290040p11.
31. Rand D, Eng JJ. Disparity between functional recovery and daily use of the upper and lower extremities during subacute stroke rehabilitation. *Neurorehabil Neural Repair*. 2012;26:76-84.
32. Wei WXJ, Fong KNK, Chung RCK, Myint J, Cheung HKY, Chow ESL. Utility of a unilateral accelerometer for monitoring upper extremity use in subacute stroke patients after discharge from hospital. *Assist Technol*. 2019;31:193-198.
33. Waddell KJ, Strube MJ, Bailey RR, et al. Does task-specific training improve upper limb performance in daily life post-stroke? *Neurorehabil Neural Repair*. 2017;31:290-300.
34. Waddell KJ, Lang CE. Comparison of self-report versus sensor-based methods for measuring the amount of upper limb activity outside the clinic. *Arch Phys Med Rehabil*. 2018;99:1913-1916.
35. Uswatte G, Miltner WH, Foo B, Varma M, Moran S, Taub E. Objective measurement of functional upper-extremity movement using accelerometer recordings transformed with a threshold filter. *Stroke*. 2000;31:662-667.
36. Uswatte G, Hobbs Qadri L. A behavioral observation system for quantifying arm activity in daily life after stroke. *Rehabil Psychol*. 2009;54:398-403.
37. Bochniewicz EM, Emmer G, McLeod A, Barth J, Dromerick AW, Lum P. Measuring functional arm movement after stroke using a single wrist-worn sensor and machine learning. *J Stroke Cerebrovasc Dis*. 2017;26:2880-2887.
38. Bleecker ML, Bolla-Wilson K, Kawas C, Agnew J. Age-specific norms for the Mini-Mental State Exam. *Neurology*. 1988;38:1565-1568.
39. Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*. 1971;9:97-113.
40. McLeod A, Bochniewicz EM, Lum PS, Holley RJ, Emmer G, Dromerick AW. Using wearable sensors and machine learning models to separate functional upper extremity use from walking-associated arm movements. *Arch Phys Med Rehabil*. 2016;97:224-231.
41. Lang CE, Wagner JM, Dromerick AW, Edwards DF. Measurement of upper-extremity function early after stroke: properties of the action research arm test. *Arch Phys Med Rehabil*. 2006;87:1605-1610.
42. Lachenbruch PA, Mickey MR. Estimation of error rates in discriminant analysis. *Technometrics*. 1968;10:1-11.
43. Pedregosa F. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
44. Peach D, Van Hoomissen J, Callender HL. Exploring the ActiLife((R)) filtration algorithm: converting raw acceleration data to counts. *Physiol Meas*. 2014;35:2359-2367.
45. Wijndaele K, Westgate K, Stephens SK, et al. Utilization and harmonization of adult accelerometry data: review and expert consensus. *Med Sci Sports Exerc*. 2015;47:2129-2139.
46. Brønd JC, Andersen LB, Arvidsson D. Generating ActiGraph counts from raw acceleration recorded by an alternative monitor. *Med Sci Sports Exerc*. 2017;49:2351-2360.
47. Bailey RR, Klaesner JW, Lang CE. Quantifying real-world upper-limb activity in nondisabled adults and adults with chronic stroke. *Neurorehabil Neural Repair*. 2015;29:969-978.
48. Lee SI, Liu X, Rajan S, Ramasarma N, Choe EK, Bonato P. A novel upper-limb function measure derived from finger-worn sensor data collected in a free-living setting. *PLoS One*. 2019;14:e0212484.
49. Friedman N, Rowe JB, Reinkensmeyer DJ, Bachman M. The manumeter: a wearable device for monitoring daily use of the wrist and fingers. *IEEE J Biomed Health Inform*. 2014;18:1804-1812.