Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the Draw a Shape Test

A P Creagh 1,2, C Simillion 2, A Scotland 2, F Lipsmeier 2, C Bernasconi 2, S Belachew 2, J van Beek 1, M Baker 2, C Gossens 3, M Lindemann 4, and M De Vos 1,3,4

1 Institute of Biomedical Engineering, University of Oxford, Oxford United Kingdom
2 F. Hoffmann-La Roche Ltd Basel, Switzerland
3 Department of Electrical Engineering and Department of Development and Regeneration, KU Leuven, Leuven, Belgium
4 Shared last authorship.

E-mail: andrew.creagh@eng.ox.ac.uk

Keywords: digital biomarkers, multiple sclerosis, hand and upper extremity function, smartphone

Supplementary material for this article is available online

Abstract

Objective: Smartphone devices may enable out-of-clinic assessments in chronic neurological diseases. We describe the Draw a Shape (DaS) Test, a smartphone-based and remotely administered test of Upper Extremity (UE) function developed for people with multiple sclerosis (PwMS). This work introduces DaS-related features that characterise UE function and impairment, and aims to demonstrate how multivariate modelling of these metrics can reliably predict the 9-Hole Peg Test (9HPT), a clinician-administered UE assessment in PwMS. Approach: The DaS Test instructed PwMS and healthy controls (HC) to trace predefined shapes on a smartphone screen. A total of 93 subjects (HC, n = 22; PwMS, n = 71) contributed both dominant and non-dominant handed DaS tests. PwMS subjects were characterised as those with normal (nPwMS, n = 50) and abnormal UE function (aPwMS, n = 21) with respect to their average 9HPT time (≤ or > 22.7 (s), respectively). L1-regularization techniques, combined with linear least squares (OLS, IRLS), or non-linear support vector (SVR) or random forest (RFR) regression were investigated as functions to map relevant DaS features to 9HPT times. Main results: It was observed that average non-dominant handed 9HPT times were more accurately predicted by DaS features ($r^2 = 0.41$, $P < 0.05$; MAE: $2.08 \pm 0.34$ (s)) than average dominant handed DaS tests. PwMS subjects were characterised as those with normal (nPwMS, n = 50) and abnormal UE function (aPwMS, n = 21) with respect to their average 9HPT time ($≤$ or $> 22.7$ (s), respectively). $L_1$-regularization techniques, combined with linear least squares (OLS, IRLS), or non-linear support vector (SVR) or random forest (RFR) regression were investigated as functions to map relevant DaS features to 9HPT times. Mean absolute error (MAE) in predicted 9HPTs was comparable to the variability of actual 9HPT times within HC, nPwMS and aPwMS groups respectively. The 9HPT however exhibited large heteroscedasticity resulting in less stable predictions of longer 9HPT times. Significance: This study demonstrates the potential of the smartphone-based DaS Test to reliably predict 9HPT times and remotely monitor UE function in PwMS.

1. Introduction

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system, affecting more than 2 million people worldwide [1]. The impairment of upper extremity (UE) function and manual dexterity resulting from sensory and motor deficits is widely reported across all subtypes of MS, although progressive MS is associated with higher prevalence of UE dysfunction and greater impairment of manual dexterity [2, 3]. UE dysfunction impacts people with MS’ (PwMS) ability to perform activities of daily living, affecting their independence, work retention and quality of life [4, 5].

While various performance tests and patient-reported outcome measures are available [6], the 9-Hole Peg Test (9HPT) is the most frequently used measure of manual dexterity in MS research, clinical trials and clinical practice [7, 8]. The 9HPT requires participants to repeatedly place and then remove nine pegs into nine holes, one at a time, as quickly as possible [7]. Performance is commonly evaluated as the time taken to
complete the task, as measured in seconds (s). Along with the Expanded Disability Status Scale (EDSS) and timed 25-foot walk (T25FW), the 9HPT is typically used as a standardized upper extremity outcome measure in MS and is integrated in the so-called Multiple Sclerosis Functional Composite (MSFC) [8–11].

The inter-rater and test-retest reliability of the 9HPT is generally high across a range of disability levels, however most studies focus on more disabled PwMS populations [7]. One large scale study with a healthy population for example reports high inter-rater reliability but only moderate test-retest reliability [12]. Regardless, the 9HPT has satisfactory discriminative and ecological validity in PwMS [13], although the coarse nature of this test and its infrequent in-clinic administration has inherent limitations. The low sampling frequency of 9HPT in-clinic administration, every 3 to 6 or 12 months in MS clinical trials or routine care practice respectively, may miss episodic manifestations of the disease due to relapses. Additionally it may not be suited for early and sensitive detection of insidious and slowly evolving change of UE function as seen in progressive MS. While the 9HPT measure is restricted to the stopwatch collection of the overall time to complete a manual dexterity task, it lacks the capacity to discriminate and quantify variable qualitative patterns of alteration of hand or finger motor skills and cannot capture intra-task fluctuations of performances. There is a need for more continuous self-administered and refined outcome measures to assess more comprehensively and reliably manual dexterity in MS and other chronic neurological disorders affecting UE function.

As smartphones become more ubiquitous worldwide in daily life, it has been proposed to utilise these devices to digitally augment specific analogue in-clinic tests. Sensor-based methods enable large quantities of objective information to be collected longitudinally and at low patient burden from both clinical and remote environments [14, 15]. These outcomes may be geared to outperform conventional rate-dependent neurological performance tests with respect to resolution, precision and sensitivity [8].

Manual dexterity assessments are easy to implement using smartphones. Recently digital spirals have been described as part of a validation study for the self-assessment of PwMS versus the MSFC, which composed of a larger smartphone-based phone application suite also testing gait, cognition and vision [16]. The drawing of shapes, and specifically the drawing of spirals (spirography), is of particular interest since it represents a direct quantitative test transformation of a commonly employed qualitative clinical method to analyse patient’s handwriting on paper [17]. Digital spirals, typically drawn using a stylus, are a common method used to analyse impairment in subjects with neurodegenerative diseases [9, 16, 18–23]. Prior efforts to develop digital drawing platforms for clinical use have been mainly focused so far on assessing patients with Parkinson’s disease (PD) through spirography [9, 18, 21, 24–29]. However, Longstaff et al (2006) assessed spiral drawing performance in PwMS [22] and Feys et al (2007) used digital circle, square and spiral drawings to quantify MS intention tremor [20]. Vianello et al (2017) have developed a smartphone-based application in healthy elderly subjects using a variety of shapes that subjects had to draw [30].

The study ‘Monitoring of Multiple Sclerosis (MS) Participants With the Use of Digital Technology (Smartphones and Smartwatches) – A Feasibility Study’ (NCT02952911) aimed at assessing the feasibility of remote patient monitoring using smartphones and smartwatches in PwMS and healthy controls (HC) [31]. In this manuscript, we focus on characterizing information to assess manual dexterity in PwMS that can be extracted from a smartphone-based ‘Draw a Shape’ (DaS) Test.

We will demonstrate how features motivated by disease pathology and UE function can be extracted from various shapes (circle, figure-8-shape, spiral, square) traced on smartphone touchscreen and how these relate to UE impairment. We will investigate further how multivariate modelling of these features can reliably predict the Nine Hole Peg test (9HPT).

2. Methods

2.1. Dataset

The NCT02952911 study aimed at assessing the feasibility of using smartphone- and smartwatch-based tests to remotely monitor PwMS and healthy controls (HC) over a 24 week period [31]. The DaS test instructed all study participants daily to draw six different shapes presented on the smartphone screen as fast and as accurate as possible, within a maximum time of 30 seconds per each attempted shape1. The six shapes to be drawn were a diagonal line bottom left to top right, a diagonal line top right to bottom left, a square, a circle, a figure-8-shape, and a spiral. The drawing had to be performed with the index finger of the tested hand, where subjects alternated each day between their dominant and non-dominant hand. Line shapes were not considered in this study. Figure 1 depicts a demonstration of the DaS test performed by a participant.

1An individual’s data or feedback on their performance outcomes, as described in this study, are not shared with the participating subjects.
Figure 1. Demonstration of the ‘Draw a Shape’ test performed by a healthy participant using a smartphone. Each participant was instructed to draw six different shapes presented on the smartphone screen as fast and as accurate as possible, within a maximum time of 30 seconds per each attempted shape: a diagonal line bottom left to top right and a diagonal line top right to bottom left (not depicted), a square, a circle, a figure-8-shape, and a spiral.

A total of 93 subjects (HC, n = 22; PwMS, n = 71) contributed both dominant and non-dominant DaS tests used for analysis in this study. Subjects were divided into normal (nPwMS) and abnormal (aPwMS) subgroups with respect to their average combined dominant and non-dominant 9HPT times over the entire study [7]. The threshold for abnormal UE function was defined by the average 9HPT times greater than the mean plus 2 standard deviations from normative data on a healthy population, pooled on dominant (9HPT threshold: 17.8 ± 2.2 (s)) and non-dominant (9HPT threshold: 18.5 ± 2.3 (s)) tests [12, 32]. Hence, the aPwMS subgroup consisted of PwMS with average 9HPT times > 22.7 (s) and the nPwMS subgroup of PwMS with average 9HPT of ≤22.7(s).

2.2. Feature Extraction

2.2.1. Raw Data Processing

Raw sensor data was collected from the smartphone touchscreen during the active DaS test and stored as x- and y-screen coordinates with a corresponding timestamp t, (x, y, t). A bespoke MATLAB script extracted attempted and completed shapes from each test, along with the corresponding hand used. All first attempts were used for further feature analysis. All data processing was performed using MATLAB vR2018a (The MathWorks, Natick, MA, USA).

2.2.2. Characterization of Manual Dexterity Performance

Multiple features were extracted from each shape capturing temporal, spatial and spatiotemporal aspects involved in the drawing task and potentially reflective of manual dexterity. Furthermore, overall test performance statistics were calculated, such as the time taken to complete all shapes and the number of shapes completed. For a full list of the features extracted please see the accompanying supplementary material. A selection of some relevant features are described below and illustrated in figures (2–5).

2.2.3. Temporal Features

Temporal features including drawing velocity, angular and radial velocities and speed distribution measures were computed to assess temporal irregularities, such as delays, smoothness, jerkiness, and rapid finger/hand movement [9, 19, 33]. The dominant frequency and power spectral density was measured for frequencies between 1–7 Hz, which was aimed to surface potential tremulous actions like that of cerebellar intention tremor or to record ataxic movements, both commonly exhibited in PwMS [32, 33]. Examples of drawing speed and power spectral density (PSD) estimate of drawing speed are illustrated in figure 2.

2.2.4. Spatial Features

Features capturing Spatial aspects of finger or hand movements were captured by features based on drawing error [9, 18, 33]. A new approach to compute drawing error is presented in this study based on a shape-matching approach known as the Hausdorff distance [34, 35]. Let X and Y be two non-empty subsets of a metric space (M, d). The Hausdorff distance \( d_H(X, Y) \) is defined as

\[
d_H(X, Y) = \max \\left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\}
\]

where sup is the supremum and inf is the infimum and distance \( d \) is computed as the Euclidean \( L_2 \) norm. This metric compares the maximum distance of one set to the nearest point in another set [36], which can be used as a basis to compute the error between the reference way-points (interpolated into a reference shape scaled to the number of pixels drawn) and the subject’s drawing attempt. The maximal Hausdorff distance is a measure of the absolute deviation from the reference shape, while the total drawing error can also be
defined as sum of the Hausdorff distances (i.e. the largest minimum distances) between the drawn and reference shape, normalized by the number of touch coordinates drawn. An example of Hausdorff distances can be found in figure 3.

2.2.5. Spatiotemporal Features

Digital drawings are unique in that they encapsulate spatial and temporal performance information simultaneously: each pixel point contains 3D data relating to the person’s hand movement at that time. This information is exploited to create discretized heat maps of touch events ($x, y, t$). A heat map not only gives a visual representation of performance but can also be used to extract a further important sub-set of features which may be sensitive to motor control or disease fluctuations.

In order to compare intensity maps for image analysis, each shape drawn is scaled to the same coordinates, while the timescales and relating colour intensities are based on global not local pixel counts. Pixels are binned into a coarser grid, both for graphical purposes and as a means to allow comparisons between subjects for each shape drawn. A graded single colour intensity scheme is used to represent pixel densities, which is transposed into an achromatic scale for image feature extraction.

Image features are extracted from both the shape drawings and their transposed heatmaps. An example of a discretised heatmap is illustrated in figure 4. Pixel intensities measure the structural composition of such images, while the drawings are also compared with an ideal drawing\(^2\) for similarities using measures such as 2D image correlation coefficient, or Mutual Information (MI) between the two images \([37]\). Image entropy, i.e. image entropy of heat map-transposed shape drawings (converted to grayscale) was calculated using

$$H = - \sum_k p_k \log_2(p_k)$$

(2)

where $k$ is the number of grey levels and $p_k$ is the probability associated with each grey level $k$. Entropy is a commonly used measure of disorder in a system and can be used to in image analysis for texture mapping \([38]\). The topography of transposed pixel intensity drawings become a function of finger movements and hence entropy a measure of smooth, non-hesitant drawing.

Further features capturing hesitation times and aspects of fine directional changes during drawing were calculated to capture elements of cognitive motor inferences known to affect UE function in MS \([4]\). Finally,

\(^2\)An ideal drawing is calculated by interpolating reference shape coordinates scaled to the same number of pixel points, irrespective the number of drawn pixels in the subject’s attempt.
a new measure, celerity, was defined by calculating the ratio of successfully passed waypoints divided by the time taken to complete the shape.

2.3. Regression Model to 9HPT

2.3.1. Data Selection

This study aims to investigate the prediction of clinical 9HPT times using features relating to UE function computed from the DaS test. Considering a simple linear regression model of the form

$$Y = \beta X^T + \mu.$$  \hspace{1cm} (3)

In this case, $X$ is the design matrix and contains the median and standard deviation of each feature per subject over all available test days for dominant and non-dominant hand tests separately. The model errors $\mu$ are assumed to be normally distributed with zero mean and constant variance, $\sigma^2$. Response variable, $Y$, is denoted as the average 9HPT time per subject over the entire study (all baseline, week 12, week 24/study completion observations considered) for each respective dominant and non-dominant handed 9HPT separately. We assume that drawing performance will generally vary depending on dominance \cite{20,32}, therefore independent models were evaluated based on dominant and non-dominant hands used.

2.3.2. Statistical Analysis

Features were assessed for non-normality by visual inspection. Those non-normal features were transformed using box-cox transformations \cite{39}. Pre-processing assessment of the response (9HPT) displayed a highly tailed distribution, and as such the 9HPT was also transformed back towards Gaussianity to help fulfill error assumptions of linear prediction \cite{40}.

Differences in median clinical metrics (EDSS, 9HPT) and feature values between subject-groups (HC, nPwMS, aPwMS) were tested using a Kruskal–Wallis test (KWt). Categorical differences in sex were investigated using a Chi-squared ($\chi^2$) test. A Brown–Forsythe test (BF) was used to evaluate the null hypothesis that the data in each categorical subject groups (HC, nPwMS, aPwMS) have equal variances, against the alternative that at least two of the data samples do not. The BF test calculates ANOVA on the absolute deviations of the data values from the group medians \cite{41}. Differences in model residuals and model prediction errors between hands and between models were also assessed using a Wilcoxon signed rank test. Pearson’s correlation ($R_p$) and Spearman’s rank correlation ($R_{sp}$) was used to assess the association of features to the 9HPT time univariately. Wilcoxon signed rank tests were used to investigate differences in 9HPT values and features between dominant and non-dominant handed tests. P-values were corrected using methods described by Benjamini and Hochberg \cite{42} in cases where multiple hypothesis testing was performed.

2.3.3. Model Generalisability

To determine the generalisability of our models, stratified 5-fold subject-wise cross-validation (CV) was employed. This consisted of randomly partitioning the dataset into $k = 5$ folds which was stratified with equal proportions of HC, nPwMS and aPwMS where possible. One set was denoted the training set (in-sample), which was further split for into smaller set for parameter selection (validation) using an internal
Figure 4. Pixel density heat map representation of spiral shapes drawn by (L-R): HC (9HPT 18.4 ± 1.2 (s)), nPwMS (9HPT 20.2 ± 2.0), aPwMS (9HPT 25.0 ± 2.1 (s)) subjects. Screen coordinates are first segmented into 2D bins of fixed width and drawing touch point coordinates are assigned to respective bins. The number of touch coordinates per bin, and hence time, are represented by heat map colour. This builds a spatio-temporal representation of digital spiral drawing which encode areas of drawing hesitation or non-movements.

5-fold CV approach. The remaining data was then denoted testing set (out-of-sample) on which predictions were made. CV was repeated 10 times with new random partitions in order to reduce bias in re-sampling and dataset splitting.

2.3.4. Model Evaluation

In order to reveal the (potentially non-linear) functional relationship between the DaS Test (as represented by the features extracted) and the associated 9HPT, a number of regression models were evaluated based on mean absolute error (MAE) and root-mean squared error (RMSE), as measured in seconds (s). MAE is defined as \( \frac{1}{N} \sum_{i=1}^{N} \left| Y_i - \hat{Y}_i \right| \); while RMSE is defined as \( \frac{1}{N} \sqrt{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2} \), where \( Y \) are the model predictions; and \( N \) are the number of observations in the training or testing set, respectively. To prevent overfitting and reduce the dimensionality (\( M \)) of the DaS features, the ‘Least absolute shrinkage and selection operator’ (LASSO) method was employed [43]. LASSO allows removal of features by shrinking some feature coefficients, \( \beta \), in our regression towards zero, filtering towards the most important measures whilst also making selection decisions on sets of collinear features. LASSO imposes the L1-norm penalty to the residual sum of squares over \( N \) test observations using non-negative values of shrinkage parameter \( \lambda \), yielding

\[
\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left( Y_i - \sum_{j=1}^{M} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{M} |\beta_j| \right\}. \tag{4}
\]

A top feature ranking table was deduced by interrogating the feature subsets selected by LASSO at each fold and repetition. The relative stability of features selected was assessed by recording the percentage of time that feature is selected at each fold and repetition.

It has been suggested that bias or prediction error can be decreased by performing a separate regression post-LASSO [44]. As such, features were selected using LASSO and those features each presented to linear models where this study investigated the performance of ordinary least squares (OLS), and iteratively re-weighted least squares (IRLS), which minimizes the weighted sum of square using a ‘bisquare weighting’ function [40, 43].

It is possible that the DaS features do not combine linearly to predict the 9HPT and such non-linear regression was also explored. Support vector regression (SVR) is a widely used technique to perform non-linear regression by mapping the feature space to a higher dimension using a ‘kernel trick’ [45]. In this case, features selected per CV-fold by LASSO are presented to SVR models, which are tuned via grid-search to determine optimal values of kernel parameter \( \gamma \), penalty parameter \( C \), and L1 soft-margin regularization parameter \( \varepsilon \). SVR models were tested using linear and Gaussian radial bias function (RBF) kernels. Further to this, non-linear random forest regression (RFR) was also investigated using the whole feature set [46]. Regressors were built on raw features and trained with a split criterion based on mean decrease in RMSE and optimised over varying numbers of trees and the number of input variables chosen at each node.
3. Results

PwMS subjects in this study were stratified into those with presumably normal (nPwMS) and abnormal (aPwMS) UE function with respect to their pooled average 9HPT times. Table 1 represents the demographic information per subject group, HC, nPwMS, aPwMS. The effects between different MS phenotypes such as primary progressive multiple sclerosis (PPMS), secondary progressive multiple sclerosis (SPMS) and relapsing remitting multiple sclerosis (RRMS) are provided in Table 1 but are not considered for analysis. The effect of differences in the male to female ratio within each subject group (HC, nPwMS and aPwMS), while imbalanced, was also not considered in subsequent analysis.

In the overall population, 9HPT times were found to be significantly different between dominant and non-dominant hands (P < 0.05). Furthermore, 9HPT times were significantly different between dominant and non-dominant hands for HC (P < 0.05) and nPwMS (P < 0.01), but not for aPwMS (P = 0.46). Two subjects’ average 9HPT times (pooled over dominant and non-dominant) were found to be 38.9 and 41.9 (s) respectively, which was greater than the mean plus 4 standard deviations (> 38.1 (s)) from the entire study population. These subjects were considered outliers with respect to the available data and were subsequently removed from final predictive analysis.

3.1. Feature Demonstration

A cross-section of relevant features are illustrated in figures (2–5). Each figure shows an example from a representative subject from each subject group: HC, nPwMS, aPwMS. Figure 1 for instance demonstrates how Hausdorff distance is calculated for the figure-8-shape, which has been observed to increase with higher 9HPT times for both dominant (R_{sp}: 0.49, P < 0.001) and non-dominant handed tests (R_{sp}: 0.51, P < 0.001).

As an example from the circle shape, drawing speed can be less smooth and more variable in nPwMS and aPwMS than HC subjects, who tend to draw faster and more consistently (figure 2). The variability in absolute drawing speed for example was significantly greater in both PwMS groups (KWT, P < 0.001) for both hands. The respective frequency distribution of drawing speed also revealed dominant peaks at multiple frequencies for more variable shape drawing.

Pixel density maps can be created based upon the relative sampling stability of the smartphone screen. The longer a finger touch pointer stays in a position the more it will be sampled, and hence a heat map representation can be built from the finger movements both temporally and spatially. Figure 4 illustrates spiral drawings represented as discretized heatmaps. Areas of hesitation and non-movement are visually apparent and characterised by dense regions of heat intensity. Image entropy encodes this accumulation of hesitation and irregularity of drawing. It was observed that higher spiral entropy values significantly correlated to higher 9HPT times (dominant, R_{sp}: 0.40, P < 0.001; non-dominant, R_{sp}: 0.45, P < 0.001). Figure 5 demonstrates the calculation of hesitation time and over shoot at the corners of square shapes. Values of drawing speed were mapped to the original drawing for visual analysis.

3.2. Feature Evaluation

This study found 311 features were significantly correlated to the 9HPT (Spearman’s rank R: P < 0.05). Wilcoxon signed rank tests between these features calculated from dominant and non-dominant handed tests revealed 70% were significantly different between handed tests (P < 0.05). It was observed that 40% of HC, 73% of nPwMS and 57% of aPwMS subject features differed significantly between hands (P < 0.05).

Table 2 describes the top ten features selected by LASSO and relative frequency that were picked for both dominant and non-dominant handed models. Image entropy was the top feature for both handed tests. However, it was computed for spiral for dominant and for the figure-8-shape for non-dominant handed
models, respectively. Among these top ten features, features extracted from the figure-8-shape and spiral shape features rank most prominently with 4/10 and 5/10 figure-8-shape features in dominant and non-dominant handed models, respectively, and 4/10 spiral features in both dominant and non-dominant handed models. Interestingly, no square-based features are represented in the top ten features for either hand. In contrast, it can be seen that spiral image entropy and figure-8-shape Hausdorff distance were picked 100% of the time in dominant handed models. figure-8-shape image entropy, the top feature for non-dominant mode was only picked 78% of the time, yet occupied the top rank. This indicates instances where this feature may not be picked at all, but when it is, it occupies regions of high importance. Nearly all top features for both handed models show moderate-to-strong Pearson’s (linear) and Spearman’s (non-linear) correlations with 9HPT (table 2).

3.3. Model Evaluation
It was observed that non-dominant handed models more accurately predicted 9HPT times (MAE: 2.08 ± 0.34 (s)) than dominant handed regression models (MAE: 2.32 ± 0.43 (s)), using simple IRLS across 5-fold CV and 10 repetitions (P < 0.01). Figure 6 compares the out-of-sample test MAE between hands as a function of number of features added to IRLS models across 5 fold CV and 10 repetitions. It can be seen that MAE decreases as more features are evaluated in both dominant and non-dominant handed models.

Non-dominant handed tests exhibited lowest MAE with 6 features (2.21 ± 0.04 (s)) compared to dominant handed tests with 16 features (1.93 ± 0.08 (s)).

Scatterplots of the raw 9HPT predictions per subject averaged over all CV-repetitions using IRLS reveal good agreement to their ground truth for dominant (r² = 0.39) and non-dominant (r² = 0.41) tests (figure 7). Breakdown of average 9HPT predictions within each subject group demonstrated that HC and nPwMS had lower MAE compared with aPwMS for both dominant (1.81 ± 1.32, 1.93 ± 1.12 (s)) and non-dominant (1.98 ± 1.15, 1.62 ± 1.10 (s)) handed models (table 3). Subject’s considered aPwMS were much more difficult to predict for both dominant (3.81 ± 2.28 (s)) and non-dominant (3.51 ± 1.56 (s)) 9HPTs.

MAE was higher in non-dominant handed models than dominant for HC subjects, whereas non-dominant handed 9HPTs were predicted more accurately than dominant for nPwMS and aPwMS. While the mean absolute error was not significantly different between hands for HC (P = 0.83) and aPwMS (P = 0.94), a larger trend was exhibited by the nPwMS group (P = 0.09). Visual corroboration between (table 3 and figure 7) reveal that at higher 9HPT values (i.e. aPwMS) the predictions were less accurate by greater magnitudes. Non-linear techniques also exhibited this pattern.

Figure 7 illustrates the intra- and inter-subject variability of the 9HPT. The within-subject variability of the 9HPT increased with higher 9HPT values (r² = 0.54, P < 0.001; R⁰.⁷³, P < 0.001; R⁰.⁷³, P < 0.001). A Brown–Forsythe test for equal variances in Y between subject groups (HC, nPwMS, aPwMS)
compares the out-of-sample test error from the four respective models built in this study. There were no significant differences observed between any of the model predictions. Finally, it has been proposed that smartphone-based tests developed for repeated assessments in remote settings may offer reliable and objective metrics that could capture a unique window in a subject's disease state and previous unseen or inappropriately estimated characteristics of MS disease [47].

4. Discussion

The present study examines UE function in PwMS with mild-to-moderate disability in comparison with HC using DaS, a self-administered digital drawing test captured on a smartphone, and demonstrates how modelling of DaS features from a test can reliably predict the average time of the clinician-administered 9-Hole Peg Test (9HPT).

It has been proposed that smartphone-based tests developed for repeated assessments in remote settings may offer reliable and objective metrics that could capture a unique window in a subject's disease state and previously unseen or inappropriately estimated characteristics of MS disease [47].

Due to the inherently heterogeneous dissemination in space and time of multiple sclerosis, PwMS experience varying levels of dysfunction or fatigue across different physical domains [48]. As a method to characterise the PwMS population in this study we have divided PwMS into those with presumed UE function abnormality (aPwMS), and those with normal UE function (nPwMS), based on average recorded 9HPT times. Abnormal 9HPT times were considered as 9HPT times greater than two standard deviations beyond hand-matched normative data from a healthy population [32]. While applying hard thresholds on clinically administered scales is a blunt stratification method, distinct attributes of each group were apparent and will be discussed with respect to features and predictions.

4.1. Feature Discussion

Previous digital upper extremity function assessments have focused mainly on the spiral drawing [9, 21, 22, 26–29] in Parkinson’s disease, while those incorporating other types of shapes or drawings have been sparse apropos the information they have extracted [30]. By considering other shapes such as the circle, square, and figure-8-shape, it was hoped to probe all aspects of hand function along with MS-specific pathological impairments such as ataxia, various tremor types, and spasticity [4, 5, 17, 49]. UE hand-motor impairments

| Shape | Feature | Description | % Chosen | $R_p$ | $R_{sp}$ |
|-------|---------|-------------|----------|-------|---------|
| **Dominant** | | | | | |
| 1 | *spiral* | IntEntropy(HM) | Image entropy | 100% | 0.48*** | 0.40*** |
| 2 | *figure-8* | HausD(X, Y) | Hausdorff distance | 100% | 0.46*** | 0.49*** |
| 3 | *circle* | nPeaksNorm(RV) | # of peaks in radial velocity | 82% | 0.38*** | 0.39*** |
| 4 | *spiral* | max($r_{peak}(x,y)$) | Maximum hesitation time | 74% | 0.43*** | 0.40*** |
| 5 | *circle* | HausD$_{middle}(x,y)$ | Hausdorff distance | 58% | –0.17 | –0.30** |
| 6 | *figure-8* | SD(AUC(x)) | Std. deviation in drawing error | 68% | 0.39*** | 0.48*** |
| 7 | *figure-8* | nPeaksNorm(v) | # of peaks in velocity | 64% | 0.43*** | 0.45*** |
| 8 | *spiral* | SD(kurt(RHO)) | Std. deviation in kurtosis in angular velocity | 52% | 0.31*** | 0.37*** |
| 9 | *figure-8* | SD(nPeaksNorm(RV)) | Std. deviation in # of peaks in radial velocity | 42% | –0.17 | –0.31** |
| 10 | *spiral* | nPeaks(RHO) | # of peaks in angular velocity | 32% | 0.39*** | 0.37*** |
| **Non-Dominant** | | | | | |
| 1 | *figure-8* | IntEntropy(HM) | Image entropy | 78% | 0.52*** | 0.47*** |
| 2 | *figure-8* | iq5(AUC(x,y)) | Interquartile range of drawing error | 96% | 0.40*** | 0.52*** |
| 3 | *spiral* | nPeaks(RV) | # of peaks in radial velocity | 84% | 0.51*** | 0.43*** |
| 4 | *figure-8* | MI(HM, HM$_{ref}$) | Spatiotemporal mutual information | 70% | –0.50*** | –0.58*** |
| 5 | *spiral* | IntEntropy(HM) | Image entropy | 66% | 0.51*** | 0.45*** |
| 6 | *spiral* | var(RHO) | Variance in angular velocity | 76% | –0.40*** | –0.41*** |
| 7 | *figure-8* | HausD(X, Y) | Hausdorff distance | 72% | 0.49*** | 0.51*** |
| 8 | *spiral* | SD(nPeaks(v)) | Std. deviation in # of peaks in drawing speed | 58% | 0.51*** | 0.37*** |
| 9 | *circle* | HD$_{x}$(HM, x, y) | Hausdorff distance × Image entropy | 52% | 0.40*** | 0.47*** |
| 10 | *figure-8* | kurt(RV) | Kurtosis in radial velocity | 36% | 0.52*** | 0.46*** |

*Table 2. Comparison of top features between dominant and non-dominant handed models as selected by lasso across 5-fold CV with 10 repetitions. Features were ranked per CV fold by increasing shrinkage regularisation parameter $\lambda$ and the percentage (%) of time that feature is chosen in the subset that minimises the CV MSE in the validation set. For a full list of the features extracted and descriptions see the supplementary material.*

*figure-8 refers to the *figure-8-shape* drawn by subjects; $R_p$, Pearson's correlation to 9HPT; $R_{sp}$, Spearman's correlation to 9HPT; $^*$ $P < 0.05$; $^*$ $P < 0.01$; $^{***} P < 0.001$.*
Figure 6. Out-of-sample test MAE (s) as features are cumulatively added to an IRLS model. Dominant and non-dominant handed models are built separately for predictions $\hat{Y}$ of average $Y$ 9HPT times (s) for each hand using 5-fold CV with 10 repetitions. Confidence intervals denote one standard deviation (SD) around the quoted mean performance across CV repetitions. Features are pre-selected and ranked within each fold using LASSO feature selection by varying shrinkage parameter $\lambda$. Minimum MAE was obtained with 16 features using the dominant handed model (MAE: 2.21 ± 0.04 (s)) and with six features using the non-dominant handed model (MAE: 1.93 ± 0.08 (s)).

manifest differently in PwMS as opposed to in Parkinson’s. This lead our study to extract a more exhaustive feature space. Both previously developed and novel features were derived and tested for their clinical validity through multivariate modelling of the 9HPT a typically used UE function test in PwMS. Figures (2–5) aim to characterise some of the DaS features developed in this study and how the level of UE impairment may influence each shape drawn and resultant feature value. Univariate analysis of these DaS features demonstrated moderate-to-strong Pearson’s and Spearman’s correlations with the 9HPT (table 2), with many coefficients comparable to a range of outcome measures for upper extremity function [7].

Consistent with our study, Feys et al [20] identified handedness as a possibly influential factor on digital drawing performance, although they were unable to test this in a healthy sub-population. Erasmus et al [32] observed a significant difference in drawing error in PwMS subgroups with cerebellar upper limb ataxia, and general worse performance in non-dominant hands across their feature set. Such differences across hands may be more amplified by MS-related impairment and in this study a greater proportion of features differed significantly between hands for nPwMS (70%) and aPwMS (57%) compared with HC (40%). Therefore, digital tests of upper extremity function that are conditioned on the hand used may be more sensitive to MS disease severity and changes in disease course.

4.2. Model Discussion
The DaS testing battery is an information-rich but dimensionally dense test. Similar features can be extracted from six different shapes, quickly accumulating the overall volume of features and contributing to redundancy. Many features exhibited collinearity within and between shapes. Hence, LASSO $L_1$-regularization was employed in order to reduce the feature space, minimise the effects of collinearity and identify important predictors of 9HPT time. Recording the relative frequency at which features were selected allows an interpretation of the feature type and shapes that are most useful to probe aspects of MS disease. Novel features such as Hausdorff distance drawing error or drawing entropy calculated from heat-map
Figure 7. Scatter plot of average 9-hole peg test (9HPT) times $Y$ versus predicted $\hat{Y}$ 9HPT times per subject for (a) dominant ($r^2 = 0.39, P < 0.05$) and (b) non-dominant ($r^2 = 0.41, P < 0.05$) handed models using IRLS. Values of $\hat{Y}$ are averaged over all CV repetitions. PwMS were grouped into presumably normal (nPwMS) versus abnormal (aPwMS) hand/arm function based on an upper limit of normal range defined as the average 9HPT time for HCs plus two standard deviations over pooled dominant and non-dominant handed tests (> 22.7 (s)). HC subjects are illustrated using green circles, nPwMS with light blue inverted triangles, and aPwMS are depicted using dark blue triangles. A black line represents perfect predictions.

Table 3. Mean absolute error (MAE) in IRLS predictions $\hat{Y}$ of average $Y$ 9HPT times (s) per subject group over dominant, non-dominant and pooled dominant and non-dominant handed models. Average MAE is calculated per subject using 5-fold CV over 10 repetitions. Standard deviations (SD) represent the SD per subject-group. Wilcoxon signed rank test between hands used for HC ($P = 0.83$), nPwMS ($P = 0.09$) and aPwMS ($P = 0.94$).

|        | Dominant | Non-Dominant | Pooled |
|--------|----------|--------------|--------|
| HC     | 1.81 ± 1.32 | 1.93 ± 1.12 | 1.97 ± 0.98 |
| nPwMS  | 1.98 ± 1.15 | 1.62 ± 1.10 | 1.90 ± 0.89 |
| aPwMS  | 3.81 ± 2.28 | 3.51 ± 1.56 | 3.75 ± 2.00 |
|        | 2.32 ± 1.66 | 2.08 ± 1.39 | 2.31 ± 1.42 |

Table 4. The out-of-sample test performance metrics of average 9-hole peg test (9HPT) times $Y$ versus predicted $\hat{Y}$ 9HPT times per dominant and non-dominant handed models. Ordinary least squares (OLS) and iteratively re-weighted least squares (IRLS) are compared to support vector (SVR) and random forest (RFR) regression. Models are built using 5-fold CV with 10 repetitions. The mean MAE and RMSE per CV is presented, where the standard deviation (SD) represents variability across CV repetitions.

|        | Dominant | Non-Dominant |
|--------|----------|--------------|
|        | MAE (s)  | RMSE (s)     | MAE (s) | RMSE (s) |
| OLS    | 2.33 ± 0.49 | 3.03 ± 0.67 | 2.09 ± 0.36 | 2.65 ± 0.42 |
| IRLS   | 2.32 ± 0.43 | 3.03 ± 0.61 | 2.08 ± 0.34 | 2.61 ± 0.39 |
| SVR    | 2.09 ± 0.42 | 2.78 ± 0.64 | 2.01 ± 0.31 | 2.54 ± 0.38 |
| RFR    | 2.09 ± 0.38 | 2.69 ± 0.62 | 1.84 ± 0.26 | 2.34 ± 0.35 |

Transformations appear as strong predictors of the 9HPT. The spiral and figure-8-shapes occupy the top feature ranks, especially for non-dominant handed tests. In addition, overall model performance (table 4, figure 6) also demonstrated that reconstruction of non-dominant 9HPT was more accurate (MAE: six features, 1.93 ± 0.08 (s)), generally with much fewer features than dominant handed 9HPTs (MAE: 16 features, 2.21 ± 0.08 (s)). Comparison of IRLS prediction error between handed models revealed that in HC, 9HPT times are more accurately predicted for dominant handed tests compared with non-dominant handed tests, whereas the improvement in prediction accuracy (i.e. reduced MAE) was strongly driven by the nPwMS and aPwMS groups (table 3). We hypothesise that more complex shapes, tested using a weaker hand, elicit a wider stratification of PwMS subjects’ performance and disease manifestation.

Furthermore, this breakdown of prediction error by subject group demonstrated that HC and nPwMS 9HPT times were accurately predicted by DaS features for both handed tests (table 3). High reconstruction
Figure 8. Scatter plot of mean 9HPT per subject, as measured in seconds (s) against the standard deviation (SD) pooled over all visits for both dominant and non-dominant handed tests and coloured by subject group. A black line represents the least squares line of best fit ($r^2 = 0.54, P < 0.001$). PwMS were grouped into presumably normal (nPwMS) versus abnormal (aPwMS) hand/arm function based on an upper limit of normal range defined as the average 9HPT time for HCs plus two standard deviations over pooled dominant and non-dominant handed tests (> 22.7 (s)). HC subjects are illustrated using green circles, nPwMS with light blue inverted triangles, and aPwMS are depicted using dark blue triangles. The 9HPT is shown to exhibit large heteroscedasticity. It was observed that the within-subject variability of the 9HPT increases with higher 9HPT values ($R_{ps}: 0.73, P < 0.001; R_{sp}: 0.57, P < 0.001$). A Brown–Forsythe test for equal variances in $Y$ between subject groups (HC, nPwMS, aPwMS) demonstrated that the between-subject variability in 9HPT also increases with each subject group ($P = 0.02$).

The accuracy of 9HPT times can be deduced considering that the MAE for HC subjects ($1.81 \pm 1.32, 1.93 \pm 1.12$ (s)) was close to the standard deviation of HC 9HPT times ($18.3 \pm 1.7, 19.1 \pm 1.8$ (s)) for both dominant and non-dominant handed tests, respectively. Similarly it was shown that MAE for nPwMS ($1.98 \pm 1.15, 1.62 \pm 1.10$ (s)) was similar to the variability of their actual 9HPT times ($19.4 \pm 2.1, 20.1 \pm 1.6$ (s)).

Higher 9HPT times (those subject observations indicated as aPwMS) were however found to be more erroneously predicted for both handed tests ($3.81 \pm 2.28, 3.51 \pm 1.56$ (s)). Visual examination of the distribution of actual 9HPT times (figure 7) demonstrated that greater 9HPT times exhibited higher variance between subject measurements, representative of an inverse Gaussian distribution. Figure 8 corroborated this heteroscedastic observation, further demonstrating that greater 9HPT times exhibited higher variance within subject measurements ($r^2 = 0.54, P < 0.001$). As 9HPT times are measured in seconds and are unbounded, hesitations, incorrect movements and the erratic impact of dropping a peg outside of the board—which are more likely in those with greater UE impairment—can compound to greater magnitudes of accumulated 9HPT times. Consequently, higher 9HPT tests become more variable and less stable, both between and within subjects with greater UE impairment. Nonetheless, the MAE for aPwMS (MAE: $3.81 \pm 2.28, 3.51 \pm 1.56$ (s)) was still less than the standard deviation of 9HPT for this group (9HPT: $26.3 \pm 5.4, 27.2 \pm 5.0$ (s)).

Comparison of the out-of-sample test prediction error across CV repetitions demonstrated that RFR and SVR, both non-linear techniques, performed slightly better than OLS or IRLS models. RFR intrinsically uses non-linear feature selection compared to SVR which is dependent on linearly selected features from LASSO, and gave minimum prediction error. However, considering the units of measurements for 9HPT is seconds (s), differences between models of 0.5 seconds can be considered minimal. As such it can be assumed that a simple linear function can accurately and adequately capture the relationship between 9HPT and DaS features.

4.3. Limitations

While the results presented in this study demonstrate the utility of using digitally captured DaS features with high concurrent validity as demonstrated by their capacity to predict clinical 9HPT times, there are a number of important limitations which need to be addressed.

First, a limitation of this work is the reliance on estimating a narrow clinical proxy of UE function as a ground truth. As discussed, the 9HPT aims to evaluate UE impairment in PwMS, but given it is measured in seconds it can exhibit a large heteroscedasticity and can be highly variable, especially for longer completion times. It was observed that higher 9HPT times had a higher variance for intra subject measurements over clinical visits (figure 8). As such, the 9HPT time should instead not be considered an exact measure of UE
impairment, but rather an estimate of a severity range of function that may be impaired. Multiple sclerosis is a heterogeneous disease which can not only manifest differently across people, but symptoms may vary within specific domains, including UE function. Some studies even suggest only moderate test-retest reliability of the 9HPT when examined in a large healthy cohort [12], rather than more disabled PwMS of other studies [7]. Reliance should therefore not be weighted on one test administered infrequently, such as the 9HPT, to effectively capture all aspects of UE function. HC and nPwMS for example were not significantly discriminative of each other based on 9HPT times (dominant: P = 0.11; non-dominant: P = 0.08). These are all limitations that should be considered when reporting predictions of any model mapping direct to the 9HPT.

The cohort analysed in this study is relatively small (n = 93 subjects). Most MS patients were mildly disabled with respect to their overall and motor specific clinical scores (table 1), and it was observed that the distribution of the 9HPT was highly tailed and skewed towards shorter 9HPT times. As a result, our models systematically underestimated 9HPT scores for aPwMS groups yet more accurately predicted shorter 9HPT times, where a greater density of similar observations were available, i.e. the HC and nPwMS subjects. The sparsity in the representation of aPwMS—who additionally are characterized by higher intra- and inter-subject 9HPT variability—limited our ability to learn a more accurate global model on longer 9HPT times. This work may therefore be biased by uneven distributions of UE impairment despite CV stratification.

Another constraint bound by low subject numbers occurs as generalisability problems across cross-validation folds. A low standard deviation in MAE and RMSE regression error (± 0.5 (s)) was observed across CV repetitions (table 4), demonstrating that results do not change with different permutations of subjects within CVs. Despite this, it was found that feature distributions may not generalise across training, validation and test sets within CV folds, leading to sub-optimal loss minimisation during the training phase. As a result, spurious feature sets and model parameters may be chosen, which can lead to more erroneous 9HPT reconstruction. Furthermore, while cross-validation itself is a popular and robust method to determine model performance and generalisability, independent test sets should ideally be used to obtain unbiased estimates of the relationship between the 9HPT and DaS features.

This study is longitudinal and data captured can span weeks’ worth of testing. A definite limitation is that the temporal aspect of this data is not fully utilised. Instead subject’s features are smoothed down as the median and standard deviation across all their available data. While the standard deviation is a coarse measure of subject variance across the study, more specific time-series modelling of the DaS Test may reveal additional detailed insights to the progression and characteristics of PwMS. For example, previous work by Prince et al [50] has shown insights into the longitudinal UE behaviour of patients with Parkinson’s Disease using a smartphone based tapping assessment.

Overall, it must be considered that NCT02952911 was a feasibility study with relatively few subjects. While this study helps establish a methodological foundation to construct models that can identify patterns of PwMS UE impairment, further studies—especially with a more heterogeneous and diverse set of subjects—and subsequent analysis will be needed to fully probe the clinical validity of remote smartphone assessments in PwMS.

5. Conclusion

This study illustrates that UE function can be assessed in remote settings using smartphone technology. The analysis from the Draw a Shape (DaS) test, a smartphone-based UE function test in which subjects trace specific shapes, expands on the feature space developed by similar studies investigating UE function in other disease areas [27–29, 51] and contextualises how new and existing features can be used to characterise UE impairment in PwMS. Multivariate modelling of these features was shown to reliably predict 9HPT times.

While perfect reconstruction of the 9HPT was not possible due to the sparsity of the dataset and the inherent limitations of the 9HPT itself, DaS features may contain a greater wealth of information supplementing beyond discrete 9HPT scores. Key advantages of digital tests like the DaS test are that they can be administered at high frequency, longitudinally and remotely in free-living environments. More frequent and ecologically valid outcome measures of UE impairment are needed to advance progressive MS research and help make clinical trials more efficient by improving power through sensitive and responsive endpoints. In this respect, the wealth of intra-task UE functional information that encapsulates the DaS feature space administered at higher and potentially daily frequency might be better suited in capturing subtle clinical changes seen in relation to the progressive course of MS than the 9HPT, which is typically administered only every 3 to 6 months. This study with ongoing further work therefore establish the foundation of how digital sensor-based assessments may enable an out-of-clinic objective augmentation of traditional rater-administered assessments of UE impairment in MS and other neurological disorders.
Acknowledgment

A P Creagh is a PhD student at the University of Oxford and acknowledges the support of F Hoffmann-La Roche Ltd; C Simillion, F Lipsmeier are employees of F Hoffmann-La Roche Ltd; C Bernasconi is a contractor for F Hoffmann-La Roche. During completion of the work related to this manuscript, S Belachew was an employee of F Hoffmann-La Roche Ltd; his current affiliation is Biogen (Cambridge, MA, USA), which was in no way involved in this work. A Scotland and M Lindemann are consultants for F Hoffmann-La Roche through Inovigilat; M Baker, J van Beek and C Gossens are employees and shareholders of F Hoffmann-La Roche Ltd; M De Vos has nothing to disclose. This research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC).

We would like to thank the following employees from F. Hoffmann-La Roche Ltd who supported and contributed to the NCT02952911 study: Atieh Bamdadian, Alessandro Barbato, Jan Beckmann, Sandro Fritz, Nicholas Pierce Heinemeier, Timothy Kilchenmann, Lito Kriara, Bernd Laub, Grégoire Pointeau, Caroline Polakowska, Marcin Puhacz, Jens Schjodt-Eriksen, Joerg Sprengel, Ralf Stubner, and Krzysztof Trybus. We would also like to acknowledge Abigail Wilson for her contribution to this manuscript.

ORCID iDs

A P Creagh https://orcid.org/0000-0002-6086-6098
C Simillion https://orcid.org/0000-0002-2026-6741
A Scotland https://orcid.org/0000-0001-9590-8617
F Lipsmeier https://orcid.org/0000-0002-8663-819X
C Bernasconi https://orcid.org/0000-0002-1019-5468
S Belachew https://orcid.org/0000-0003-3976-1950
J van Beek https://orcid.org/0000-0002-5641-0004
M Baker https://orcid.org/0000-0002-6854-0375
C Gossens https://orcid.org/0000-0003-2660-5895
M Lindemann https://orcid.org/0000-0003-1010-3890
M De Vos https://orcid.org/0000-0002-3482-5145

References

[1] Goldenberg M M 2012 Multiple sclerosis review Pharmacy Ther. 37 175
[2] Holper L, Coenen M, Weise A, Stucki G, Cieza A and Kesselring J 2010 Characterization of functioning in multiple sclerosis using the ICF J. Neurol. 257 103–13
[3] Poole J L, Nakamoto T, McNulty T, Montoya J R, Weill D, Dieruf K and Skipper B 2010 Dexterity, visual perception and activities of daily living in persons with multiple sclerosis Occup. Ther. Health Care. 24 159–70
[4] Learmonth Y C, Plutti L A and Molf R W 2015 Generalised cognitive motor interference in multiple sclerosis Gait and posture 42 96–100
[5] Yozbatiran N, Baskurt F, Baskurt Z, Ozakbas S and Idiman E 2006 Motor assessment of upper extremity function and its relation with fatigue, cognitive function and quality of life in multiple sclerosis patients J. Neurol. Sci. 246 117–22
[6] Lamers I and Feyes P 2014 Assessing upper limb function in multiple sclerosis Mult. Scler. 20 775–84
[7] Feyes P, Lamers I, Francis G, Benedict R, Phillips G, LaRocca N, Hudson L D, Rudick R and 2017 Multiple Sclerosis Outcome Assessments Consortium. The nine-hole peg test as a manual dexterity performance measure for multiple sclerosis Mult. Scler. J. 23 711–20
[8] Cutter G R et al 1999 Development of a multiple sclerosis functional composite as a clinical trial outcome measure Brain 122 871–82
[9] Memedi M, Sadikov A, Groznik V, Žabkar J, Možina M, Johansson A, Haubenberger D and Nyholm D 2015 Automatic spiral analysis for objective assessment of motor symptoms in Parkinson's disease Sensors (Basel, Switzerland) 15 23727–44
[10] Rudick R A, Cutter G and Reingold S 2002 The multiple sclerosis functional composite: a new clinical outcome measure for multiple sclerosis trials Mult. Scler. J. 8 359–65
[11] Cadavid D et al 2017 The EDDS-plus, an improved endpoint for disability progression in secondary progressive multiple sclerosis Mult. Scler. 23 94–105
[12] Grice K O, Vogel K A, Viet L, Mitchell A, Muniz S and Vollmer M A 2003 Adult norms for a commercially available nine hole peg test for finger dexterity Am. J. Occup. Ther. 57 570–3
[13] Kragt J, van der Linden F A H, Nielsen J M, Uitdehaag B M J and Polman C H 2006 Clinical impact of 20% worsening on timed 25-foot walk and 9-hole peg test in multiple sclerosis Mult. Scler. J. 12 594–8
[14] Bove R et al 2015 Evaluating more naturalistic outcome measures: A 1-year smartphone study in multiple sclerosis Neurol Neuroimmunol Neuroinflamm 2 e162
[15] Dean J M and Silverman M 2015 The utilization of smartphone devices to enhance clinical interventions Movement Disorders 30 5463
[16] Maillart E et al 2010 Mscopilot, a new multiple sclerosis self-assessment digital solution: results of a comparative study versus standard tests European J. Neurol. 2019 429–36
[17] Alusi S H, Worthington J, Glickman S, Findley L J and Bain P G 2000 Evaluation of three different ways of assessing tremor in multiple sclerosis J. Neurol. Neurosurg. Psychiatry 68 756–60
[18] Aghanavesi S, Nyholm D, Senek M, Bergquist F and Memedi M 2017 A smartphone-based system to quantify dexterity in Parkinson's disease patients Informatics in Medicine Unlocked 9 11–17

[19] Banaszkiewicz K, Rudzinska M, Bukowczan S, Iwowski A and Szczudlik A 2009 Spiral drawing time as a measure of bradykinesia Neurol. Neurochir. Pol. 43 16–21

[20] Feys P, Helsen W, Prinsmel A, Isbrouck S, Wang S and Liu X 2007 Digitised spirography as an evaluation tool for intention tremor in multiple sclerosis J. Neurol. Methods 160 309–16

[21] Wang M, Wang B, Zou J and Nakamura M 2012 A new quantitative evaluation method of spiral drawing for patients with Parkinson's disease based on a polar coordinate system with varying origin Physica A: Stat. Mech. Appl. 391 4377–88

[22] Longstaff M G and Heath R A 2006 Spiral drawing performance as an indicator of fine motor function in people with multiple sclerosis Hum. Mov. Sci. 25 474–91

[23] Wang H, Qiping Y, Kurtis M M, Floyd A G, Smith W A and Pullman S J 2008 Spiral analysis—improved clinical utility with center detection J. Neurol. Methods 171 264–70

[24] Michael P C, Hans-Leo T, Vincent Filoteo J, David S and James B I 2006 Quantitative measurement of handwriting in the assessment of drug-induced Parkinsonism Hum. Mov. Sci. 25 510–22

[25] Xuguang L, Camille B C, Shou-Yan W, Zajicek J, and Peter G B 2005 Quantifying drug-induced dyskinesias in the arms using digitised spiral-drawing tasks. J. Neurosci. Methods 144 47–52

[26] Galli M, Vimercati S L, Manetti E, Cinolimon V, Albertini G and De Pandis M F Spiral analysis in subjects with Parkinson's disease before and after levodopa treatment: a new protocol with stereophotogrammetric systems J. Appl. Biomech. Funct. Mater. 12 2014

[27] Memedi M, Aghanavesi S and Westin J 2016 A method for measuring Parkinson's disease related temporal irregularity in spiral drawings 2016 IEEE-EMBS Int. Conf. on Biomedical and Health Informatics (BHI) pp 410–13

[28] Sadikov A, Groznik V, Možina M, Žabkar J, Nyholm D, Memedi M, and Georgiev D 2017 Feasibility of spirography features for objective assessment of motor function in Parkinson's disease. Artif. Intell. Medi. 81 54–62

[29] Somayeh A, Memedi M, Dougherty M, Nyholm D and Westin J 2017 Measuring temporal irregularity in spiral drawings of patients with Parkinson's disease 21st Int. Congress of Parkinson's Disease and Movement Disorders (Vancouver, BC, Canada) vol 32, pp s252

[30] Vianello A, Chittaro L, Burigat S and Budai R 2017 Motorbrain: A mobile app for the assessment of users' motor performance in neurology Comput. Methods Programs Biomed. 143 35–47

[31] Midaglia L et al 2019 Adherence and satisfaction of smartphone-and smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: Nonrandomized interventional feasibility study J. Med. Internet Res. 21 e14863

[32] Erasmus L-P, Sarno S, Albrecht H, Schwecht M, Pöllmann W and Nicolais K 2001 Measurement of atactic symptoms with a graphic tablet: standard values in controls and validity in multiple sclerosis patients J. Neurosci. Methods 108 25–37

[33] Feys P, Helsen W, Prinsmel A, Isbrouck S, Wang S and Liu X 2007 Digitised spirography as an evaluation tool for intention tremor in multiple sclerosis J. Neurol. Methods 160 309–16

[34] Dubuisson M-P and Jain A K 1994 A modified hausdorff distance for object matching, pattern recognition, 1994 Vol. 1-Conf. A: Computer Vision and Image Processing., Proc. of the 12th IAPR Int. Conf. on IEEE (Piscataway, NJ, IEEE) vol 1 pp 566–8

[35] Daniel P H, Klanderman G A and Rucklidge W J 1993 Comparing images using the hausdorff distance IEEE Trans. Pattern Anal. Mach. Intell. 15 850–63

[36] Veltkamp R C Shape matching: similarity measures and algorithms Proc. Int. Conf. on Shape Modeling and Applications pp 188–97

[37] Plum J P W, Antoine Maintz J B and Viergever M A 2003 Mutual-information-based registration of medical images: a survey Med. Image Anal. 7 19–41

[38] Tsai D-Y, Lee Y and Matsuyma E 2008 Information entropy measure for evaluation of image quality Journal of Digital Imaging 21 338–47

[39] Sakia R M 1992 The box-cox transformation technique: a review The Statistician 41 169–78

[40] Clark V, Dunn O J and Mickey R M 1974 Applied Statistics, Analysis of Variance and Regression (New York: Wiley)

[41] Brown M B and Forsythe A B 1974 The small sample behavior of some statistics which test the equality of several means Technometrics 16 129–32

[42] Benjamini Y and Hochberg Y 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing J. Royal Stat. Soc.: Series B (Methodological) 57 289–300

[43] Tibshirani R 1996 Regression shrinkage and selection via the lasso J. Royal Stat. Soc. Series B (Methodological) 58 267–88

[44] Belloni A and Chernozhukov V 2013 Least squares after model selection in high-dimensional sparse models Bernoulli 19 521–47

[45] Scholkopf B, Smola A J and Bach F 2002 Learning With Kernels: Support Vector Machines, Regularization, Optimization and Beyond (Cambridge, MA: MIT Press)

[46] Breiman L, Friedman J, Stone C J and Olshen R A 1984 Classification and Regression Trees (Boca Raton, FL: CRC)

[47] Memedi M, Nyholm D, Johansson A, Paalhagen S, Willows T, Widner H, Linder J and Westin J 2015 Validity and responsiveness of at-home touch screen assessments in advanced Parkinson's disease IEEE J. Biomed. Health Inform. 19 1829–34

[48] Compton A and Coles A 2002 Multiple sclerosis Lancet 359 1221–31

[49] Alusi S H, Worthington J, Glickman S and Bain P G 2001 A study of tremor in multiple sclerosis Brain 124 720–30

[50] Prince J, Arora S and Maarten de V 2018 Big data in Parkinson’s disease: using smartphones to remotely detect longitudinal disease Journal of Neurology, Neurosurgery and Psychiatry 89 1699–711

[51] Poonam Zham D K Kumar P D, Arjunan S P and Raghav S 2017 Distinguishing different stages of Parkinson's disease using composite index of speed and pen-pressure of sketching a spiral Frontiers Neurol. 8 435