Physiol. Meas.

Smartphone-Based Remote Assessment of Upper Extremity Function for Multiple Sclerosis Using the FLOODLIGHT 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[2], M. Baker[2], C. Gossens[2], M. Lindemann[2‡] and M. De Vos[1‡]

[1] Institute of Biomedical Engineering, University of Oxford, UK;
[2] F. Hoffmann-La Roche Ltd., Basel, CH;
[‡] Shared last authorship.

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

Abstract. *Objective*: Smartphone devices may enable out-of-clinic assessments in chronic neurological diseases. We describe the FLOODLIGHT 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 FLOODLIGHT 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 ($\leq$ 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.
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 9HPTs (r^2=0.39, P <0.05; MAE: 2.32 $\pm$ 0.43 [s]), using simple linear IRLS (P <0.01). Moreover, it was found that the 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.

Keywords: Digital Biomarkers; Multiple Sclerosis; Hand and Upper Extremity Function; Smartphone

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, 9, 10, 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 rater-dependent neurological performance tests with respect to resolution, precision and sensitivity [8].

Manual dexterity assessments are easy to implement using smartphones. Recently

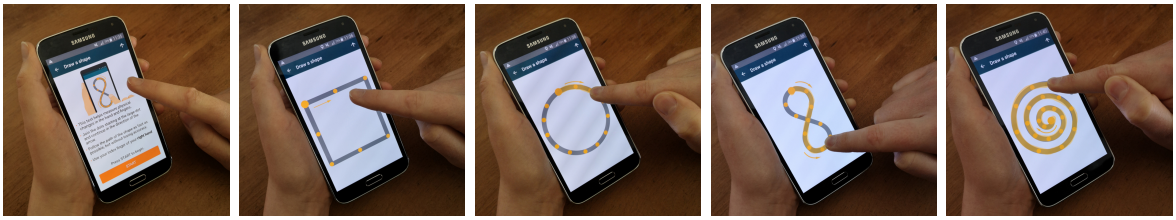


Figure 1: Demonstration of the FLOODLIGHT "Draw a Shape" test performed by a healthy participant using a smartphone. Each participant was instructed to draw six different shapes presented on the smartphone screen as fast and as accurate as possible, within a maximum time of 30 seconds per each attempted 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*.

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 [18, 19, 20, 21, 22, 16, 9, 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 [18, 24, 21, 25, 26, 27, 9, 28, 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 FLOODLIGHT study (NCT02952911) was a proof-of-concept study to assess 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 FLOODLIGHT's "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

FLOODLIGHT was a 24 week, proof-of-concept study aimed at assessing the feasibility of using smartphone- and smartwatch-based tests to remotely monitor PwMS and healthy controls (HC) [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 shape‡. 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. 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 the mean plus 2 standard deviations from normative data on a healthy population, pooled on dominant (9HPT threshold: 17.8 + 2(2.2) [s]) and non-dominant (9HPT threshold: 18.5 + 2(2.3) [s]) tests [32, 12]. 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].

‡ An individual's data or feedback on their performance outcomes, as described in this study, are not shared with the participating subjects.

Table 1: Demographics and characteristics for Healthy Controls (HC) and PwMS subgroups, stated as mean ± SD, over the entire study where applicable;

	HC (n=22)	nPwMS (n=50)	aPwMS (n=21)	$P^{(a)}$
# tests per subject	74 ± 59	114 ± 51	107 ± 56	$*^1$
Age, year	34 ± 9	40 ± 8	40 ± 8	n.s.[1]
Dominant hand (Right/Left)	(19/3)	(44/6)	(20/1)	n.s.[2]
Male/Female	15/7	14/36	8/13	$**^2$
MS diagnosis, (PPMS/SPMS/RRMS)	NA	2/1/47	1/3/17	n.s.[2]
EDSS	NA	2.1 ± 1.26	3.3 ± 1.4	$***^1$
9HPT, seconds				
(dominant)	18.3 ± 1.7	19.4 ± 2.1	26.3 ± 5.4	$***^1$
(non-dominant)	19.1 ± 1.8	20.1 ± 1.6	27.2 ± 5.0	$***^1$
$P^{(b)}$	$*^3$	$**^3$	n.s.[3]	

nPwMS: PwMS with average 9HPT ≤ 22.7 [s]; aPwMS: PwMS with average 9HPT >22.7 [s]; PPMS, Primary Progressive Multiple Sclerosis; SPMS, Secondary Progressive Multiple Sclerosis; RRMS, Relapse Remitting Multiple Sclerosis; EDSS, Expanded Disability Status Scale; 9HPT, 9-Hole Peg Test; NA, not applicable.

[(a)] P-value between groups [(b)] P-value between hands

[1] Kruskal-Wallis by ranks tests the null hypothesis that the continuous data in each categorical group (HC, nPwMS, aPwMS) comes from the same distribution;

[2] Chi-squared (χ^2) tests differences between categorical groups (HC, nPwMS, aPwMS);

[3] Wilcoxon signed rank tests 9HPT values between hands used for each group;

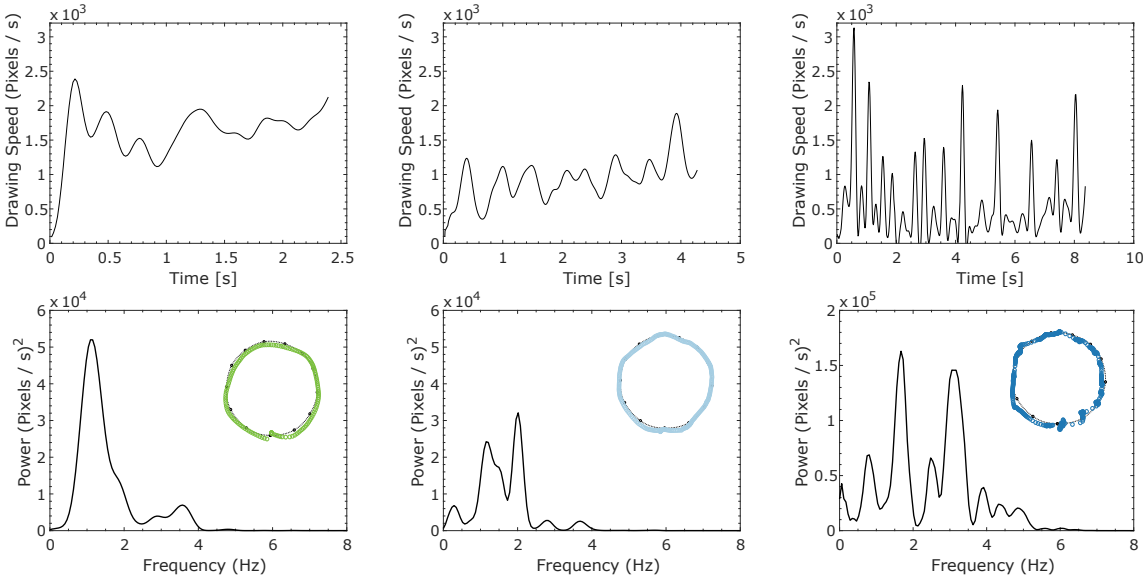
* P <0.05; ** P <0.01; *** P <0.001, n.s. - not significant.

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 [33, 34, 35]. 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 [34, 32]. Examples of drawing speed and power spectral density (PSD) estimate of drawing speed are illustrated in Figure 1.

2.2.4. Spatial Features Features capturing Spatial aspects of finger or hand movements were captured by features based on drawing error [18, 34, 35]. A new approach to compute drawing error is presented in this study based on a shape-matching approach known as the Hausdorff distance [36, 37]. 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:

$$d_H(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right\} \tag{2.1}$$

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



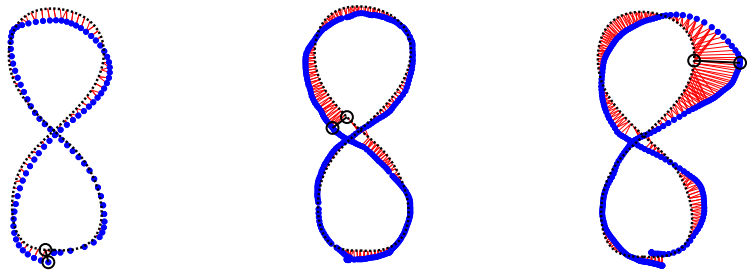
(a) HC (b) nPwMS (c) aPwMS

Figure 2: Example illustrations of *circle* shape 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. Red points depict actual pixel points drawn relative to interpolated reference coordinates (black). The top row demonstrates examples of drawing speed for duration of time to draw each respective shape. Time series speed signal was first filtered using a low pass filter with a cut off frequency of 8 Hz. The bottom row represents the power spectral density (PSD) estimate of drawing speed which was computed using a Hamming window. Note the time and PSD axis scale between the figures.

nearest point in another set [38], 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 3-D data relating to the persons 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



(a) HC (b) nPwMS (c) aPwMs

Figure 3: Examples of *figure-8-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. Figure depicts actual pixel points drawn (blue) relative to interpolated reference coordinates (black). Hausdorff Distance query points are illustrated with red lines and maximal Hausdorff Distances ($HausD$, as measured in Pixels) are highlighted with black circles; HC (64 Pixels) , nPwMS (90 Pixels), aPwMS (229 Pixels). The total drawing error ($HausDError$) 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. In this example he normalised drawing errors are: HC (31 Pixels), nPwMS (38 Pixels), aPwMS (68 Pixels).

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.

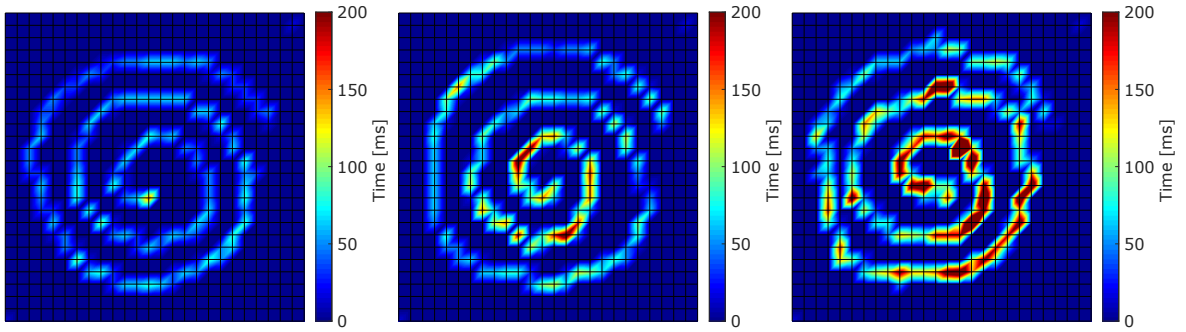
Image features are extracted from both the shape drawings and their transposed heatmaps. An example of a discretised heatmap is illustrated in Figure 4. Pixel intensities measure the structural composition of such images, while the drawings are also compared with an ideal drawing§ for similarities using measures such as 2-D image correlation coefficient, or Mutual Information (MI) between the two images [39]. Image entropy, i.e. image entropy of heat map-transposed shape drawings (converted to grayscale) was calculated using:

$$H = -\sum_{k} p_k \log_2(p_k) \tag{2.2}$$

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 [40]. 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, a new measure, celerity, was defined by calculating

§ 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) HC (b) nPwMS (c) aPwMS

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.

the ratio of successfully passed waypoints divided by the time taken to complete the shape.

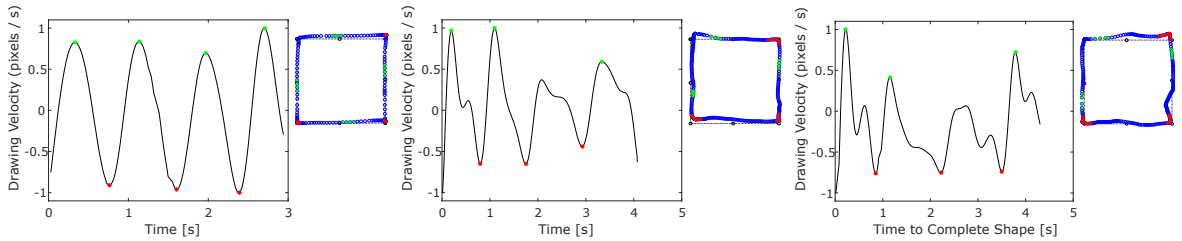
2.3. Regression Model to 9HPT

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:

$$Y = \beta X^{\intercal} + \mu \tag{2.3}$$

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 μ are assumed to be normally distributed with zero mean and constant variance, σ^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 [32, 20], 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 [41]. 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 fulfil error assumptions of linear prediction [42].
Differences in median clinical metrics (EDSS, 9HPT) and feature values between



(a) HC (b) nPwMS (c) aPwMS

Figure 5: Example of *square* 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. Normalised drawing velocity is shown with local maximum and minimum speed highlighted in green and red respectively. Corresponding points on shape drawing are also illustrated where local minimum velocity points represent dwell (hesitation) time at corner locations. Note the time axis between figures.

subject-groups (HC, nPwMS, aPwMS) were tested using a Kruskal-Wallis test (KWt). Categorical differences in sex were investigated using a Chi-squared (χ^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 [43]. 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_{ps}) 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 [44] 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 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 nonlinear) 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 absolute error (RMSE). MAE is defined as: $\frac{1}{N}\sum_{i=1}^{N}|Y_i - \hat{Y}_i|$; while RMSE is defined as $\frac{1}{N}\sqrt{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}$, where $\hat{Y}$ 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 [45]. LASSO allows removal of features by shrinking some feature coefficients, β, in our regression towards zero, filtering towards the most important measures whilst also making selection decisions on sets of collinear features. LASSO imposes the L_1-norm penalty to the residual sum of squares over N test observations using non-negative values of shrinkage parameter λ, yielding:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(Y_i - \sum_{j=1}^{M} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{M} |\beta_j| \right\} \tag{2.4}$$

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 [46]. 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 [42] [45].

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" [47]. 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 γ, penalty parameter C, and L_1 soft-margin regularization parameter ϵ. 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 [48]. 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 pultiple 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 FLOODLIGHT 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 3 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 4 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 models 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 (r^2=0.39) and non-dominant (r^2=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

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 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.

	Shape	Feature	Description	% Chosen	R_{ps}	R_{sp}
			Dominant			
1	*spiral*	$ImEntropy(HM)$	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*	$\max(t_{pixel}(x,y))$	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*	$ImEntropy(HM)$	Image entropy	78%	0.52***	0.47***
2	*figure-8*	$iqr(AUC(x,y))$	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*	$ImEntropy(HM)$	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*	$HD_E(HM,x,y)$	Hausdorff distance x Image entropy	52%	0.40***	0.47***
10	*figure-8*	$kurt(RV)$	Kurtosis in radial velocity	36%	0.52***	0.46***

figure-8 referes to the *figure-8-shape* drawn by subjects;
R_{ps}, Pearson's correlation to 9HPT; R_{sp}, Spearman's correlation to 9HPT;
* P <0.05; ** P <0.01; *** P <0.001.

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 (r^2=0.54, P <0.001; 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 increased with each subject group (P =0.02). In

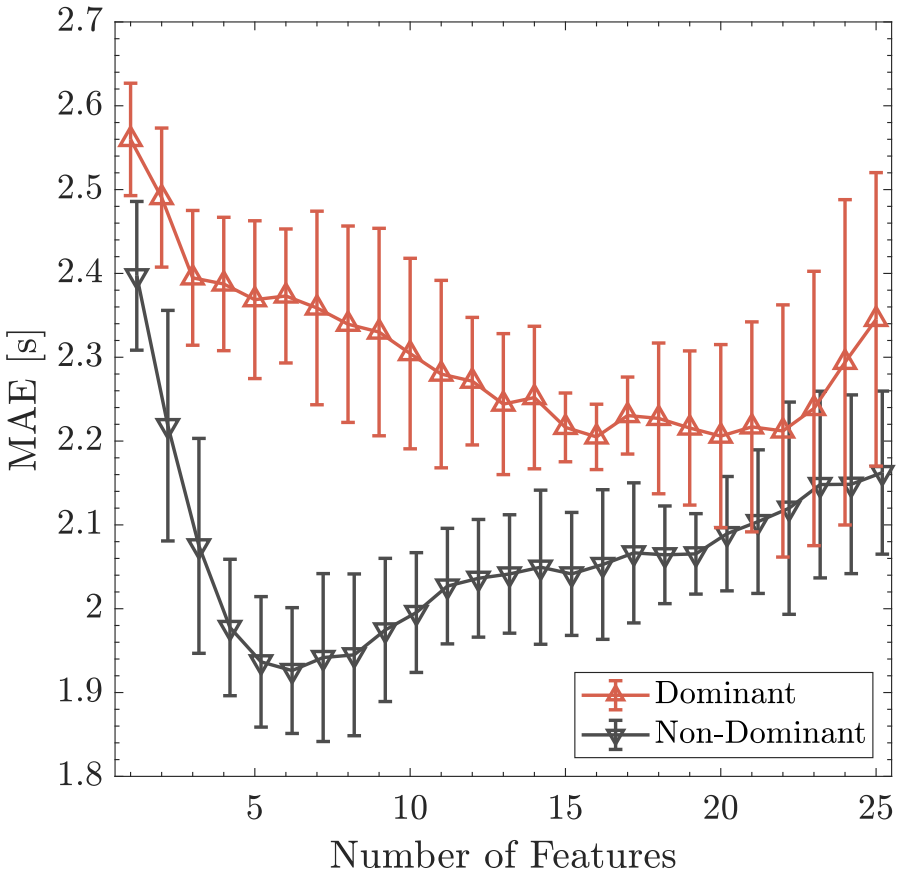


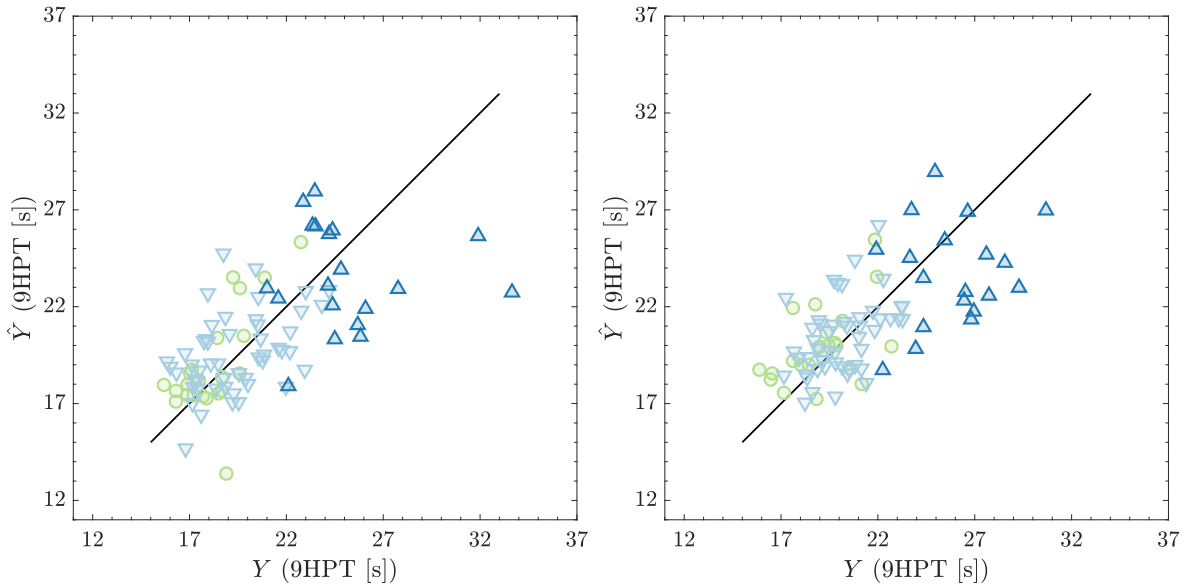
Figure 6: Out-of-sample test MAE [s] as features are cumulatively added to an IRLS model. Dominant and non-dominant handed models are built separately for predictions $\hat{Y}$ of average Y 9HPT times [s] for each hand using 5-fold CV with 10 repetitions. Confidence intervals denote one standard deviation (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 λ. Minimum MAE was obtained with 16 features using the dominant handed model (MAE: 2.21 $\pm$ 0.04 [s]) and with 6 features using the non-dominant handed model (MAE: 1.93 $\pm$ 0.08 [s]).

concordance with Table 3, aPwMS were shown to exhibit greater variability than HC and nPwMS, where higher 9HPT times tend to have much greater within- and between-subject variance. Finally, Table 4 compares the out-of-sample test error from the 4 respective models built in this study. There were no significant differences observed between any of the model predictions.

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



(a) Dominant (b) Non-Dominant

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.

window in a subject's disease state and previously unseen or inappropriately estimated characteristics of MS disease [49].

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 [50]. As a method to characterise the PwMS population in this study we have divided PwMS into those with presumed UE function abnormality (aPwMS),

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

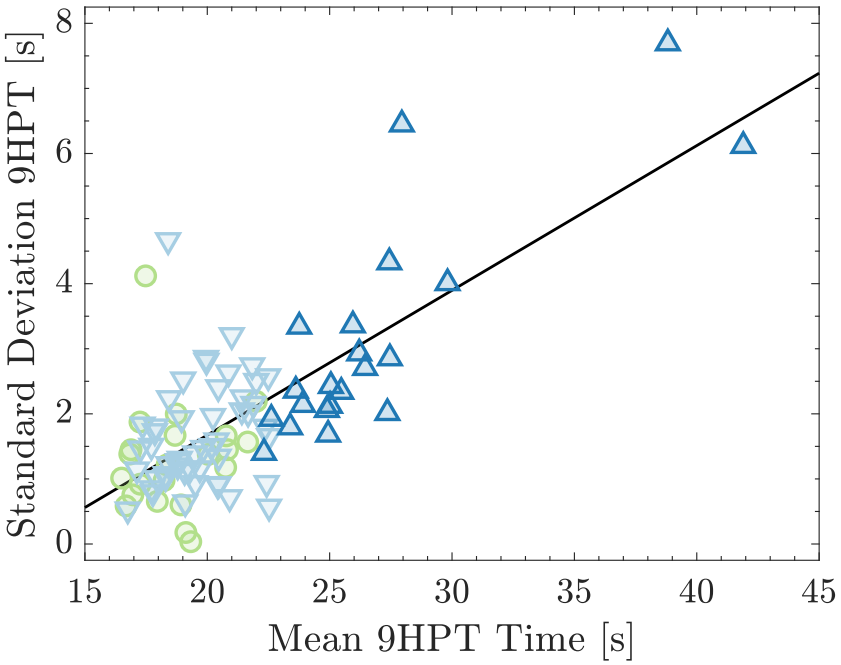


Figure 8: Scatter plot of mean 9HPT per subject, as measured in seconds [s] against the standard deviation (SD) pooled over all visits for both dominant and non-dominant handed tests and coloured by subject group. A black line represents the least squares line of best fit (r^2=0.54, P <0.001). PwMS were 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).

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

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 [26, 27, 9, 28, 29, 21, 22] 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 [17, 51, 4, 5]. UEHMF impairments 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, collated by Feys *et al* [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 6 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 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: 6 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 accuracy of 9HPT times can be deduced considering that the MAE for HC subjects (1.81 ± 1.32, 1.93 ± 1.12 [s]) was close to the standard deviation of HC 9HPT times (18.3 ± 1.7, 19.1 ± 1.8 [s]) for both dominant and non-dominant handed tests, respectively. Similarly it was shown that MAE for nPwMS (1.98 ± 1.15, 1.62 ± 1.10 [s]) was similar to the variability of their actual 9HPT times (19.4 ± 2.1, 20.1 ± 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 ± 2.28, 3.51 ± 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 ± 2.28, 3.51 ± 1.56 [s]) was still less than the standard deviation of 9HPT for this group (9HPT: 26.3 ± 5.4, 27.2 ± 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 FLOODLIGHT 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 ($\pm$ 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 mean 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* [52] 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 FLOODLIGHT was a proof-of-concept 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, 28, 29, 53] 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.

Acknowledgments

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 Inovigate; 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 FLOODLIGHT 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, Jöerg Sprengel, Ralf Stubner, and Krzysztof Trybus. We would also like to acknowledge Abigail Wilson for her contribution to this manuscript.

References

[1] Marvin M Goldenberg. Multiple sclerosis review. *Pharmacy and Therapeutics*, 37(3):175, 2012.

[2] L. Holper, M. Coenen, A. Weise, G. Stucki, A. Cieza, and J. Kesselring. Characterization of functioning in multiple sclerosis using the icf. *J Neurol*, 257(1):103–13, 2010.

[3] J. L. Poole, T. Nakamoto, T. McNulty, J. R. Montoya, D. Weill, K. Dieruf, and B. Skipper. Dexterity, visual perception, and activities of daily living in persons with multiple sclerosis. *Occup Ther Health Care*, 24(2):159–70, 2010.

[4] YC Learmonth, LA Pilutti, and RW Motl. Generalised cognitive motor interference in multiple sclerosis. *Gait posture*, 42(1):96–100, 2015.

[5] N. Yozbatiran, F. Baskurt, Z. Baskurt, S. Ozakbas, and E. Idiman. 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(1-2):117–22, 2006.

[6] I. Lamers and P. Feys. Assessing upper limb function in multiple sclerosis. *Mult Scler*, 20(7):775–84, 2014.

[7] Peter Feys, Ilse Lamers, Gordon Francis, Ralph Benedict, Glenn Phillips, Nicholas LaRocca, Lynn D Hudson, Richard Rudick, and Multiple Sclerosis Outcome Assessments Consortium. The nine-hole peg test as a manual dexterity performance measure for multiple sclerosis. *Multiple Sclerosis Journal*, 23(5):711–720, 2017.

[8] G. R. Cutter, M. L. Baier, R. A. Rudick, D. L. Cookfair, J. S. Fischer, J. Petkau, K. Syndulko, B. G. Weinshenker, J. P. Antel, C. Confavreux, G. W. Ellison, F. Lublin, A. E. Miller, S. M. Rao, S. Reingold, A. Thompson, and E. Willoughby. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. *Brain*, 122 (Pt 5):871–82, 1999.

[9] Mevludin Memedi, Aleksander Sadikov, Vida Groznik, Jure Žabkar, Martin Možina, Filip Bergquist, Anders Johansson, Dietrich Haubenberger, and Dag Nyholm. Automatic spiral analysis for objective assessment of motor symptoms in parkinson's disease. *Sensors (Basel, Switzerland)*, 15(9):23727–23744, 2015.

[10] RA Rudick, G Cutter, and S Reingold. The multiple sclerosis functional composite: a new clinical outcome measure for multiple sclerosis trials. *Multiple Sclerosis Journal*, 8(5):359–365, 2002.

[11] D. Cadavid, J. A. Cohen, M. S. Freedman, M. D. Goldman, H. P. Hartung, E. Havrdova, D. Jeffery, R. Kapoor, A. Miller, F. Sellebjerg, D. Kinch, S. Lee, S. Shang, and D. Mikol. The edss-plus, an improved endpoint for disability progression in secondary progressive multiple sclerosis. *Mult Scler*, 23(1):94–105, 2017.

[12] Kimatha Oxford Grice, Kimberly A Vogel, Viet Le, Ana Mitchell, Sonia Muniz, and Mary Ann Vollmer. Adult norms for a commercially available nine hole peg test for finger dexterity. *American Journal of Occupational Therapy*, 57(5):570–573, 2003.

[13] JJ Kragt, F AH van der Linden, JM Nielsen, B MJ Uitdehaag, and CH Polman. Clinical impact of 20% worsening on timed 25-foot walk and 9-hole peg test in multiple sclerosis. *Multiple Sclerosis Journal*, 12(5):594–598, 2006.

[14] R. Bove, C. C. White, G. Giovannoni, B. Glanz, V. Golubchikov, J. Hujol, C. Jennings, D. Langdon, M. Lee, A. Legedza, J. Paskavitz, S. Prasad, J. Richert, A. Robbins, S. Roberts, H. Weiner, R. Ramachandran, M. Botfield, and P. L. De Jager. Evaluating more naturalistic outcome measures: A 1-year smartphone study in multiple sclerosis. *Neurol Neuroimmunol Neuroinflamm*, 2(6):e162, 2015.

[15] J.M. Dean and M. Silverman. The utilization of smartphone devices to enhance clinical interventions. *Movement Disorders*, 30:S463, 2015.

[16] Elisabeth Maillart, Pierre Labauge, Mikael Cohen, Adil Maarouf, Sandra Vukusic, Cécile Donzé, Philippe Gallien, Jérôme De Sèze, Bertrand Bourre, and Thibault Moreau. Mscopilot, a new multiple sclerosis self-assessment digital solution: results of a comparative study versus standard tests. *European journal of neurology*, 2019.

[17] SH Alusi, J Worthington, S Glickman, LJ Findley, and PG Bain. Evaluation of three different ways of assessing tremor in multiple sclerosis. *Journal of Neurology, Neurosurgery Psychiatry*, 68(6):756–760, 2000.

[18] Somayeh Aghanavesi, Dag Nyholm, Marina Senek, Filip Bergquist, and Mevludin Memedi. A smartphone-based system to quantify dexterity in parkinson's disease patients. *Informatics in Medicine Unlocked*, 9:11–17, 2017.

[19] K. Banaszkiewicz, M. Rudzinska, S. Bukowczan, A. Izworski, and A. Szczudlik. Spiral drawing time as a measure of bradykinesia. *Neurol Neurochir Pol*, 43(1):16–21, 2009.

[20] P. Feys, W. Helsen, A. Prinsmel, S. Ilsbroukx, S. Wang, and X. Liu. Digitised spirography as an evaluation tool for intention tremor in multiple sclerosis. *J Neurosci Methods*, 160(2):309–16, 2007.

[21] Min Wang, Bei Wang, Junzhong Zou, and Masatoshi Nakamura. 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: Statistical Mechanics and its Applications*, 391(18):4377–4388, 2012.

[22] Mitchell Grant Longstaff and Richard A Heath. Spiral drawing performance as an indicator of fine motor function in people with multiple sclerosis. *Human movement science*, 25(4):474–491, 2006.

[23] Hongzhi Wang, Qiping Yu, Mónica M. Kurtis, Alicia G. Floyd, Whitney A. Smith, and Seth L. Pullman. Spiral analysis—improved clinical utility with center detection. *Journal of Neuroscience Methods*, 171(2):264–270, 2008.

[24] Michael P. Caligiuri, Hans-Leo Teulings, J. Vincent Filoteo, David Song, and James B. Lohr. Quantitative measurement of handwriting in the assessment of drug-induced parkinsonism. *Human Movement Science*, 25(4–5):510–522, 2006.

[25] Xuguang Liu, Camille B. Carroll, Shou-Yan Wang, John Zajicek, and Peter G. Bain. Quantifying drug-induced dyskinesias in the arms using digitised spiral-drawing tasks. *Journal of Neuroscience Methods*, 144(1):47–52, 2005.

[26] Manuela Galli, Sara L Vimercati, Elena Manetti, Veronica Cimolin, Giorgio Albertini, and Maria F De Pandis. Spiral analysis in subjects with parkinson's disease before and after levodopa treatment: a new protocol with stereophotogrammetric systems. *Journal of applied biomaterials functional materials*, 12(2), 2014.

[27] M. Memedi, S. Aghanavesi, and J. Westin. A method for measuring parkinson's disease related temporal irregularity in spiral drawings. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 410–413, 2016.

[28] Aleksander Sadikov, Vida Groznik, Martin Možina, Jure Žabkar, Dag Nyholm, Mevludin Memedi, and Dejan Georgiev. Feasibility of spirography features for objective assessment of motor function in parkinson's disease. *Artificial Intelligence in Medicine*, 2017.

[29] Aghanavesi Somayeh, Mevludin Memedi, Mark Dougherty, Dag Nyholm, and Jerker Westin. Measuring temporal irregularity in spiral drawings of patients with parkinson's disease. In *21st International Congress of Parkinson's Disease and Movement Disorders, Vancouver, BC, Canada*, volume 32, 2017.

[30] Andrea Vianello, Luca Chittaro, Stefano Burigat, and Riccardo Budai. Motorbrain: A mobile app for the assessment of users' motor performance in neurology. *computer methods and programs in biomedicine*, 143:35–47, 2017.

[31] Luciana Midaglia, Patricia Mulero, Xavier Montalban, Jennifer Graves, Stephen L Hauser, Laura Julian, Michael Baker, Jan Schadrack, Christian Gossens, and Alf Scotland. Adherence and satisfaction of smartphone-and smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: Nonrandomized interventional feasibility study. *Journal of medical Internet research*, 21(8):e14863, 2019.

[32] Lutz-Peter Erasmus, Stefania Sarno, Holger Albrecht, Martina Schwecht, Walter Pöllmann, and Nicolaus König. Measurement of ataxic symptoms with a graphic tablet: standard values in controls and validity in multiple sclerosis patients. *Journal of Neuroscience Methods*, 108(1):25–37, 2001.

[33] Krzysztof Banaszkiewicz, Monika Rudzińska, Sylwia Bukowczan, Andrzej Izworski, and Andrzej Szczudlik. Spiral drawing time as a measure of bradykinesia. *Neurologia i neurochirurgia polska*, 43(1):16–21, 2009.

[34] Peter Feys, Werner Helsen, Ann Prinsmel, Stephan Ilsbroukx, Shouyan Wang, and

Xuguang Liu. Digitised spirography as an evaluation tool for intention tremor in multiple sclerosis. *Journal of Neuroscience Methods*, 160(2):309–316, 2007.

[35] M. Memedi, A. Sadikov, V. Groznik, J. Zabkar, M. Mozina, F. Bergquist, A. Johansson, D. Haubenberger, and D. Nyholm. Automatic spiral analysis for objective assessment of motor symptoms in parkinson's disease. *Sensors (Basel)*, 15(9):23727–44, 2015.

[36] M-P Dubuisson and Anil K Jain. A modified hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 566–568. IEEE, 1994.

[37] Daniel P. Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.

[38] R. C. Veltkamp. Shape matching: similarity measures and algorithms. In *Proceedings International Conference on Shape Modeling and Applications*, pages 188–197.

[39] Josien PW Pluim, JB Antoine Maintz, and Max A Viergever. Mutual-information-based registration of medical images: a survey. *IEEE transactions on medical imaging*, 22(8):986–1004, 2003.

[40] Du-Yih Tsai, Yongbum Lee, and Eri Matsuyama. Information entropy measure for evaluation of image quality. *Journal of Digital Imaging*, 21(3):338–347, 2008.

[41] RM Sakia. The box-cox transformation technique: a review. *The statistician*, pages 169–178, 1992.

[42] Virginia Clark, OJ Dunn, and RM Mickey. *Applied statistics, analysis of variance and regression*. Wiley, 1974.

[43] Morton B Brown and Alan B Forsythe. The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 16(1):129–132, 1974.

[44] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[45] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[46] Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[47] Bernhard Schölkopf, Alexander J Smola, and Francis Bach. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[48] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[49] M. Memedi, D. Nyholm, A. Johansson, S. Pålhagen, T. Willows, H. Widner, J. Linder, and J. Westin. Validity and responsiveness of at-home touch screen

assessments in advanced parkinsonx0027;s disease. *IEEE Journal of Biomedical and Health Informatics*, 19(6):1829–1834, 2015.

[50] A. Compston and A. Coles. Multiple sclerosis. *Lancet*, 359(9313):1221–31, 2002.

[51] Sundus Husni Alusi, J Worthington, S Glickman, and PG Bain. A study of tremor in multiple sclerosis. *Brain*, 124(4):720–730, 2001.

[52] John Prince, Siddharth Arora, and Maarten de Vos. Big data in parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes. *Physiological measurement*, 39(4):044005, 2018.

[53] Poonam Zham, Dinesh K. Kumar, Peter Dabnichki, Sridhar Poosapadi Arjunan, and Sanjay Raghav. Distinguishing different stages of parkinson's disease using composite index of speed and pen-pressure of sketching a spiral. *Frontiers in Neurology*, 8(435), 2017.