Longitudinal Trend Monitoring of Multiple Sclerosis Ambulation using Smartphones

Andrew P. Creagh*[1], Frank Dondelinger[2], Florian Lipsmeier[2], Michael. Lindemann†[2] and Maarten De Vos†[3,4]

***Abstract*—*Goal*: Smartphone and wearable devices may act as powerful tools to remotely monitor physical function in people with neurodegenerative and autoimmune diseases from out-of-clinic environments. Detection of progression onset or worsening of symptoms is especially important in people living with multiple sclerosis (PwMS) in order to enable optimally adapted therapeutic strategies. MS is a disease whose symptoms typically follow subtle and fluctuating disease courses, patient-to-patient, and over time. Current in-clinic assessments are often too infrequently administered to reflect longitudinal changes in MS impairment that impact daily life. This work, therefore, explores how smartphones can administer daily two-minute walking assessments to monitor PwMS physical function at home. *Methods*: Remotely collected smartphone inertial sensor data was transformed through state-of-the-art Deep Convolutional Neural Networks, to estimate a participant's daily ambulatory-related disease severity, longitudinally over a 24-week study. *Results*: This study demonstrated that smartphone-based ambulatory severity outcomes could accurately estimate MS level of disability, as measured by the EDSS score (r^2: 0.56,p <0.001). Furthermore, longitudinal severity outcomes were shown to accurately reflect individual participants' level of disability over the study duration. *Conclusion*: Smartphone-based assessments, that can be performed by patients from their home environments, could greatly augment standard in-clinic outcomes for neurodegenerative diseases. The ability to understand the impact of disease on daily-life between clinical visits, through objective digital outcomes, paves the way forward to better measure and identify signs of disease progression that may be occurring out-of-clinic, to monitor how different patients respond to various treatments, and to ultimately enable the development of better, and more personalised care.**

***Index Terms*—Gait, deep learning, multiple sclerosis, digital biomarkers, smartphones**

I. Introduction

Neurodegenerative diseases, such as multiple sclerosis (MS), frequently fluctuate over time, and patient-to-patient, ensuring that it is notoriously difficult to quantify effective therapeutic interventions and disease management techniques. Current in-clinic assessments are often too infrequent to track changes in MS impairment over time. Importantly, it has been shown that earlier identification of changes in PwMS impairment are important to identify and provide better therapeutic strategies [1]. As a result, there exists a great opportunity to augment current clinical examination strategies, to integrate methods that accurately and remotely monitor disease-related changes and deterioration, that may occur at home and between clinician visits.

[1]Institute of Biomedical Engineering, University of Oxford, UK; [2]F. Hoffmann-La Roche Ltd, Basel, CH; [3]Department of Electrical Engineering [4]Department of Development and Regeneration, KU Leuven, BE;
*Corresponding author e-mail: (andrew.creagh@eng.ox.ac.uk);
†These authors jointly supervised.

Although MS follows a highly heterogeneous and subject-specific disease course, the disease profiles can be grouped into four clinical phenotypes which are based on disease progression [2], [3]: the majority of PwMS will initially experience Relapsing–remitting MS (RRMS), a state dominated by sudden acute symptoms developing (a "relapse") over days before generally plateauing over weeks or months [4], termed "remission". RRMS generally affects 85% of PwMS and disease activity typically occurs acutely at a sub-clinical level. Secondary-progressive MS (SPMS) can occur in some RRMS patients, where the disease course continues to worsen with or without periods of remission. Half of RRMS patients will go onto develop SPMS [5]–[7]. Those experiencing consistent but worsening symptoms can be thought of as having Primary-progressive MS (PPMS) [4], [5], [7] (roughly 10% of PwMS [6]). Progressive-relapsing MS is more rare (affecting fewer than 5% of PwMS); it occurs from diagnoses as a progressive disease course, with periods of relapse, but without any remission periods.

Digital smartphone-based assessments offer the ability to objectively monitor disability levels in people with multiple sclerosis (PwMS) from out-of-clinic, at home environments [8]–[12]. For instance, smartphone-based monitoring was exemplified in a recent investigation by Bove *et al.* (2015) [13], with this study demonstrating the feasibility of administering daily smartphone-based tasks to PwMS over a one-year period. These technologies can provide new data-driven metrics for clinical decision-making during in-clinic visits [14] and may be more accurate than conventional clinical outcomes, recorded at infrequent visits, to detect subtle, progressive, sub-clinical changes or trends in long-term PwMS disability [13].

Alterations during ambulation (gait) due to MS are a amongst the most common indication of MS impairment [17]–[22]. It has been shown that gait impairment affects quality of life, health status and productivity [23] in persons with MS (PwMS), with the prevalence of these reported impairments between 75% and 90% [24]. PwMS can display postural instability [18], gait variability [19]–[21] and fatigue [22] during various stages of disease progression. The gold-standard assessment of overall disability in MS is the Expanded Disability Status Scale (EDSS) [25], however there are specific functional domain assessments such as the Timed 25-Foot Walk (T25FW), which is part of the Multiple Sclerosis Functional Composite score [26], [27], and the Two-Minute Walk Test (2MWT) which also assesses physical gait function and fatigue in PwMS [28]. In recent years however, there has been a shift towards the adoption of body worn sensors to objectively evaluate ambulatory performance in PwMS, circumventing the need for resource-intensive and

TABLE I: Population Demographics[1]. Clinical scores taken as the average per subject over the entire study, where the mean ± standard deviation across population are reported; RRMS, Relapsing-remitting MS; PPMS, Primary-progressive MS; SPMS, Secondary-progressive MS; EDSS, Expanded Disability Status Scale; T25FW, the Timed 25-Foot Walk; EDSS (amb.) refers to the ambulation sub-score as part of the EDSS; [s], indicates measurement in seconds;

	HC (n=24)	PwMSmild[a] (n=52)	PwMSmod[b] (n=21)
Age	35.6 ± 8.9	39.3 ± 8.3	40.5 ± 6.9
Sex (M/F)	18/6	16/36	7/14
RRMS/PPMS/SPMS		52/0/0	14/3/4
EDSS		1.7 ± 0.8	4.2 ± 0.7
EDSS (amb.)		0.1 ± 0.3	1.9 ± 1.5
T25FW [s]	5.0 ± 0.9	5.3 ± 0.9	7.9 ± 2.2

[1] For more information on the study population we refer the reader to [15] and [16];
[a] PwMS with average EDSS < 3.5; [b] PwMS with average EDSS ≥ 3.5;

expensive gait laboratory equipment, but also opening up the possibility to measure physical function outside of standard clinical settings [14], [20], [21], [29]–[34]

This study builds upon our previous investigations [12], [15], [35], where we have shown how inertial sensors contained within consumer-based smartphones can be used to characterise gait impairments in PwMS from a remotely administered Two-Minute Walk Test (2MWT). The latter study first introduced how state-of-the-art Deep Convolutional Neural Networks (DCNN) can be applied to remote 2MWT smartphone sensor data to determine a study participants' status: such as healthy, PwMS with mild, or PwMS with moderate disability. The work presented here aims to evaluate how these DCNN severity predictions from daily 2MWTs can characterise the status of healthy participants versus PwMS with mild, or PwMS with moderate disability over a 24 week period.

II. Methods

A. Data

The FLOODLIGHT (FL) proof-of-concept (PoC) app was trialled in a 24-week, prospective study in PwMS and HCs (NCT02952911) to assess the feasibility of remote patient monitoring using smartphone (and smartwatch) devices [11], [16]. Participants were provided with a preconfigured smartphone (Samsung Galaxy S7) and smartwatch (Motorola 360 Sport) with the Floodlight PoC app installed. A total of 97 participants (24 HC subjects; 52 mildly disabled, PwMSmild, EDSS [0-3]; 21 moderately disabled PwMSmod, EDSS [3.5-5.5]) contributed data which was recorded from a 2MWT performed out-of-clinic [15]. Subjects were requested to perform a 2MWT daily over a 24-week period, and were clinically assessed at baseline, week 12 and week 24. For further information on the FL app, dataset, and population demographics we direct the reader to [16] and specifically to our previous work [12], [15], which this study expands upon. Table I depicts the population demographics for this study. All participants provided informed consent, and the ethical approval was obtained from ethics committee of the Hospital Universitari Vall d'Hebron, Barcelona, Spain and the institutional review board of the University of California San Francisco, San Francisco, CA, USA, prior to study initiation.

B. Estimating Ambulatory-related Disease Severity from Smartphone Sensor Data

Smartphone inertial sensor data was recorded while participants performed a daily, at home, two minute walk test (2MWT). The raw accelerometer sensor data from each 2MWT were then partitioned into multiple vector sequences (epochs), of 2.56 sec (128 samples/epoch) with 50% overlap between adjacent windows. A Deep Convolutional Neural Network (DCNN) was then trained to classify a given epoch as having been performed by a HC, PwMSmild or PwMSmod participant. The DCNN model implemented has previously been introduced in [12], where the network was first pre-trained on the UCI smartphone-based Human Activity Recognition (HAR) dataset, and thereafter fine-tuned on the data in FL for MS severity classification. Briefly, a DCNN applied a series of one-dimensional kernels on the raw sensor epoch $\mathbf{x}_n$ with an input (channel 1-4): $\mathbf{X}_n = (\boldsymbol{a}_x, \boldsymbol{a}_y, \boldsymbol{a}_z, \|\boldsymbol{a}\|)^\top$, where $\boldsymbol{a}$ are acceleration vectors for the x-, y- and z- components containing samples $\boldsymbol{a} = (x_1, x_2, ..., x_T)$ and $\|\boldsymbol{a}\|$ refers to original orientation invariant signal magnitude. The DCNN consisted of four causal convolutional blocks with batch normalisation (BN) layers ($momentum = 0.99$, $\epsilon = 1e^{-2}$): the 1st block extracted 32 fixed filters with a width of 9 samples, stride length of 1 (9×1), with l_2-norm regularisation ($\lambda = 1e^{-3}$); the 2nd and 3rd blocks learned 64 filters, with width (3×1); the 4th block learned 128 filters with a width of 6 (6×1), followed by a final 3-class dense fully connected softmax layer. Max pooling operations were also applied in the 2nd and 4th layers with pool size p=2 and down-scaled by stride factor s=2. Smartphone orientation augmentation was performed randomly rotating sensor-channel axis during training [36]. The DCNN was trained to minimise a multi-class categorical cross-entropy loss function for $k \in \{hc, mild, mod\}$ to learn the optimal network weights $\mathbf{w}$, using an *Adam* optimization algorithm with a learning rate $lr = 1e-5$, as well as $\beta_1 = 0.9$ and $\beta_2 = 0.999$ which determined the exponential decay rates for the moment estimates of the gradient [37], [38]. The network outputs are interpreted as $\hat{y}_k(\mathbf{x}_n, \mathbf{w}) = p(y_k = k|\mathbf{x}_n)$. As such, $\hat{y}_k$ can be thought of as the probability that a given epoch $\mathbf{x}_n$ belonged to class k. A continuous estimate of severity,

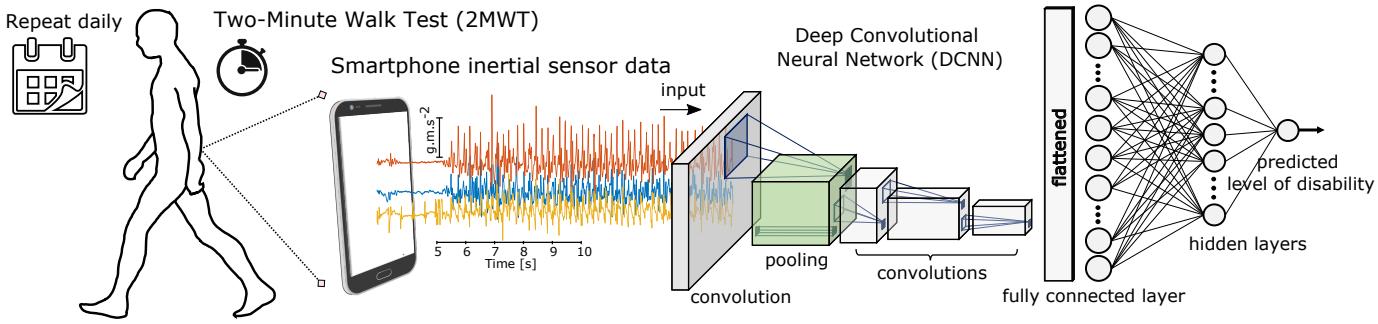


Fig. 1: **Demonstration of how deep learning algorithms can transform smartphone measurements to predict MS patient severity symptoms between clinical visits.** Illustration of Deep Convolutional Neural Network (DCNN) applied to raw smartphone inertial sensor data collected from a remotely executed Two-Minute Walk Test (2MWT), performed daily for 24-weeks.

the predicted level of MS disability, can then be captured by taking an average of all epoch predictions over a test for a given assessment day, d such that:

$$\hat{y}_d = \frac{1}{N}\sum_{n=1}^{N} \arg\max_k (p(\hat{y}_k = k|\mathbf{x}_n)) \tag{1}$$

where N are the number of windowed epochs for a given test date, d, and k lies in an ordinal range of $[0, 1, ...K]$. Therefore $\hat{y}_d$ will be continuous such that $0 \leq \hat{y}_d \leq K$ and can conceptualised as a naïve estimate of MS disease severity, mapping a predicted level of disability ranging from healthy to mild to moderate.

Models were trained using a stratified, subject-wise, 5-fold cross-validation (CV), with subjects randomly partitioned into one of k=5 folds, as described previously in [15]. One set was denoted the training set (in-sample), which was further split into a smaller set for validation, using roughly 10% of the training subjects. Predictions were evaluated on all available 2MWTS per subject in each of the (out-of-sample) test sets.

C. Longitudinal Trend Monitoring of Remote Smartphone-Based Outcomes

Longitudinal trends of specific participants were examined as a time-series by considering the severity estimates $\hat{y}$ of repeated 2MWTs over all their available data for the duration of the FL study. While participants were requested to perform a daily Two-Minute Walk Test (2MWT), some test-dates may be missing; it was also observed that various participants had differing adherence rates during the study. The number valid 2MWT recordings contributed for each subject group over the study duration is presented in appendix figure A.1. Further information related to participant adherence in the study is reported previously in [11], [16]. As the goal of this work was to perform longitudinal analysis of participants severity, namely visualise the average severity trends over time, missing 2MWT outcomes were first imputed using piecewise linear interpolation (PLI) [39], by considering $\hat{y}$ as a time-series to impute missing test severity observations on a given date. Note: imputed 2MWTs were only included for calculation of average trend estimation for individual participants and not for model evaluation. Next, a simple trend estimation was applied to the time sequence of severity estimates ($\hat{y}$) across days (d) using a $d-$ centred linear moving average filter (MAF).

$$\hat{z}[i] = \frac{1}{2N+1}\sum_{j=0}^{2N} \hat{y}[i+N-j] \tag{2}$$

where $\hat{y}[\cdot]$ is the input sequence (severity estimates) and $\hat{z}[\cdot]$ is the output (filtered) sequence (moving severity estimate) for each d^{th} day; $2N+1$ defines the order of the filter, in this case the number of days d used in the moving average. A 7-day window was implemented in order to capture the trends in $\hat{y}_d$ over the study duration.

D. Statistical Analysis

The association between estimated continuous disease severity and EDSS was tested using (linear) Pearson's (r) and (non-linear) Spearman's (ρ) correlation coefficient. A non-parametric Kruskal-Wallis (KWt) test by ranks assessed the median severity estimate between HC, PwMSmild, and PwMSmod groups. Statistical differences in smartphone severity estimates were also investigated within participants over the duration of the study. For instance, mean differences in severity estimates before and after specific events, such as the reporting of a relapse, were assessed using a t-test. In cases where severity estimates had unequal variances before/after an event, as determined by a Brown-Forsythe (BF) test by medians [40], a Welch's t-test correction was applied. Furthermore, differences in median severity estimates before/after each event were also assessed with a non-parametric Mann-Whitney U test.

III. Results

A. Digitally Estimated Severity Outcome

A continuous disease severity outcome was created by averaging all 2MWT predictions (i.e. HC, PwMSmild, PwMSmod) for each participant, calculated from each of the out-of-sample test sets during cross-validation. A disease severity outcome therefore mapped a posterior probability ranging from healthy, to mild, and to moderate for each subject. The distribution of the average severity per subject was displayed in figure 2, and demonstrated the positive relationship between average severity outcome and average EDSS per participant (over all

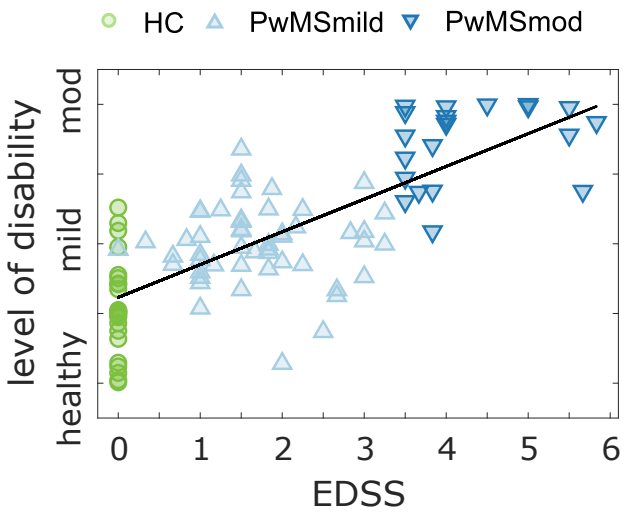


Fig. 2: **The relationship between the continuous disease severity outcome estimate, EDSS and subject group**. Figure depicts the scatter plot demonstrating the positive correlation (r : 0.75; ρ : 0.71; $p < 0.001$) between the average severity outcome and average EDSS score per subject. A DCNN model was constructed based on the average class predictions (HC, PwMSmild, PwMSmod) per subject over all 2MWTs, creating an estimated continuous prediction probability distribution, ranging from healthy to moderate MS. Each point therefore represents the average estimated severity outcome (probability) for that subject. A black line represents the line of best fit between severity and EDSS (r^2 : $0.56, p < 0.001$).

available EDSS scores for that participant), (Pearson's r : 0.75; Spearman's ρ : 0.71; $p < 0.001$, r^2 : $0.56, p < 0.001$). Model classification performance can also be determined by thresholding the estimated continuous level of disability in figure 2, at the boundaries between HC, PwMSmild, and PwMSgroups, as reported in [12].

B. Longitudinal Characterisation of Digitally Estimated Severity Outcomes

Disease severity outcomes were evaluated for each 2MWT performed per subject. As a result, longitudinal trends in ambulatory impairment can be monitored by examining daily 2MWT estimates for participants over the duration of the FL study. While the 24-week duration of the study and relatively low level of baseline impairment of the participants meant that we did not observe meaningful progression at the study cohort level, we could still investigate the ability of our methodology to capture participant-specific longitudinal trends. For example, figure 3 examines the longitudinal severity estimate outcome for for various representative correctly classified HC, PwMSmild and PwMSmod participants. Individual 2MWT ambulatory severity estimates are depicted from the 0^{th} week until study completion in week 24, where dashed black lines represented site-visits where participants were assessed clinically. Blue lines depicted the 7-day average trend in severity outcomes.

Figure 3a first depicted a HC subject. This participant was examined at baseline (week 0; EDSS 0, T25FW: 5 [s]), midway through the study (week 12; EDSS 0[1]; T25FW: 4.5 [s]) and at the study completion (week 24; EDSS: 0; T25FW: N/A[2] [s]). It was observed that this subject was predicted as healthy with a low severity, consistently across the entire study. Many variations in severity outcomes were smoothed out across the 7-day moving average. Similarly, figure 3b demonstrated a correctly classified, stable, PwMSmild participant over the duration of the study. This participant was also clinically examined at week 0 (EDSS: 2.5; T25FW: 6.8) week 12 (EDSS: 2.5; T25FW: 6.5 [s]) and at week 24 (EDSS: 3; T25FW: 6.6). In comparison, figure 3c demonstrated a stable PwMSmod subject. This participant was examined at baseline (week 0; EDSS 3.5, T25FW: 5.4 [s]) and midway through the study[3] (week 12; EDSS 4.5; T25FW: 4.9 [s]).

During the FLOODLIGHT study, four PwMS subjects reported relapses using the FLOODLIGHT application on their smartphone during the study. These participants' ambulatory-based 2MWT severity estimates were investigated in figure 4.

Figure 4a depicts the longitudinal severity outcome trend for a PwMSmild subject who reported a relapse during the FL PoC study. A black line depicts the date of relapse on-setting during week 3, which was recorded by the participant using the FLOODLIGHT application on their smartphone. Dashed black lines represent site-visits where the participant was assessed clinically. This subject was examined at baseline (week 0; EDSS 1.5, T25FW: 4.9 [s]), week 12 (EDSS 1.5; T25FW: N/A[4]) and at the study completion in week 25 (EDSS: N/A[5]; T25FW: 5.5 [s]). In week 4, 7 days after reporting a relapse, the participant was assessed during an "unscheduled visit" where they exhibited a worsening of MS symptoms, i.e. an increase in EDSS and gait related T25FW (EDSS: 2.5; T25FW: 7.5 [s]). Their relapse was evaluated as a spinal topography outbreak.

Figure 4b assesses the severity outcomes for another PwMSmild subject, with a clinical examination at baseline (week 0; EDSS: 1.5, T25FW: 4.9 [s]), and during visit 2 (week 12; EDSS: 1.5, T25FW: 6 [s]). This participant reported a relapse during week 23 where their EDSS rose by +1 during their clinical examination during study completion (EDSS: 2.5, T25FW: 5.9 [s]).

Figure 4c examines the longitudinal severity outcome trend for a PwMSmod subject who reported a relapse during the FL PoC study. A black line depicts the date of relapse on-setting during week 13. This subject was clinically examined at baseline (week 0; EDSS 3.5, T25FW: 4.9 [s]), midway through the study (week 12; EDSS 3.5; T25FW: 6.6 [s]) and at the study completion (week 24; EDSS: 3.5; T25FW: 10.3 [s]).

Lastly, the ambulatory severity estimates for a PwMSmod participant who self-reported a relapse is shown in figure 4d. This participant's clinical examination was reported at baseline (week 0; EDSS: 3.5; T25FW: 7.8 [s]), mid-study as (week 12; EDSS: 4; T25FW: 10.5 [s]) and during study completion as (week 24; EDSS: 4; T25FW: 11.5 [s]). During the clinical examination in week 12, this participant also reported non-relapse adverse clinical events, occurring on unspecified dates sometime between weeks 8 and 12. As such, the time between

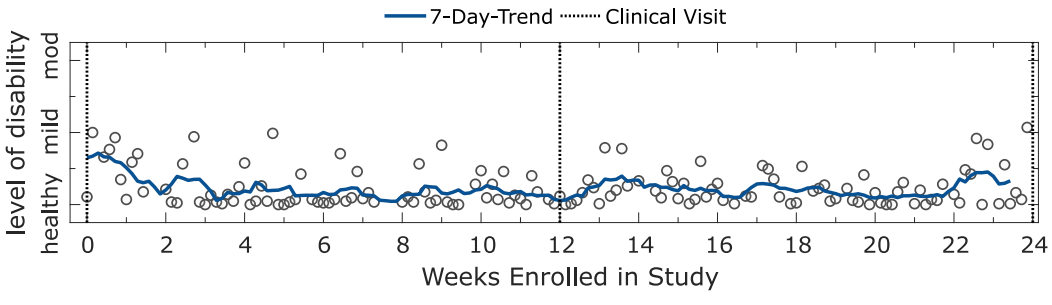
[1]Note: an EDSS of zero in this case refers to a normal neurological exam, the subject is healthy and has no disability.

[2] N/A denotes scores not assessed at this visit.

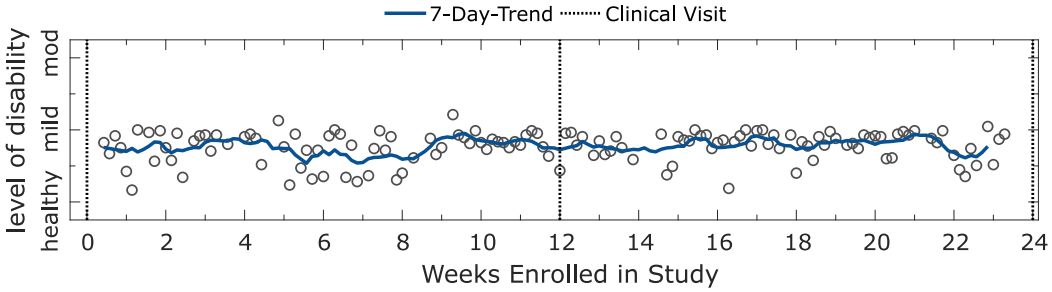
[3]Note: clinical assessment scores were not made available for this participant at study completion.

[4]See footnote 2

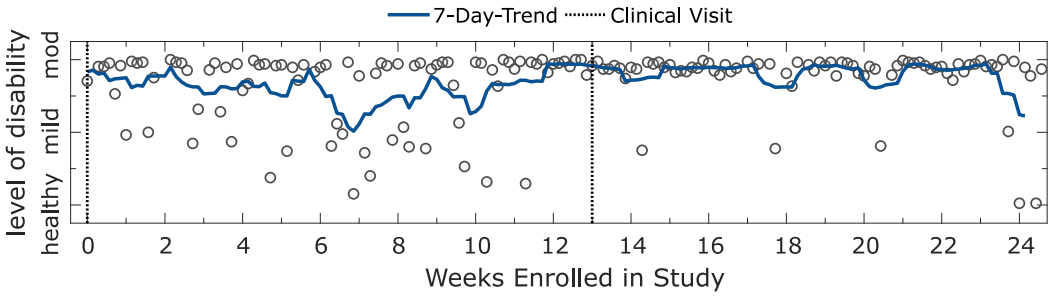
[5]See footnote 2



(a) Correctly Classified HC Subject (T25FW: 4.8±0.4 [s])



(b) Correctly Classified PwMSmild Subject (EDSS: 2.7±0.3; T25FW: 6.6±0.2 [s])



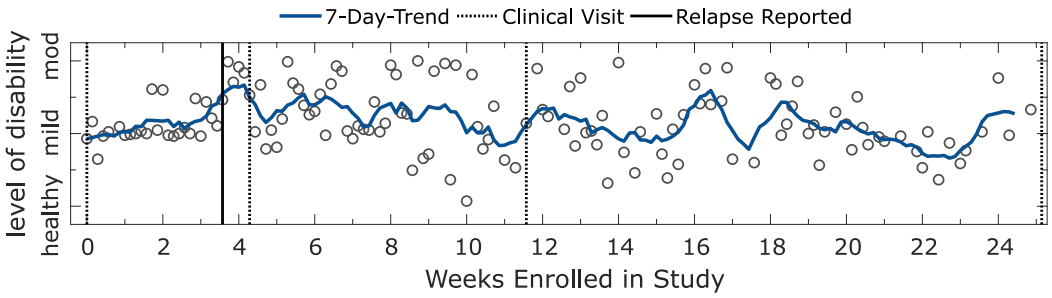
(c) Correctly Classified PwMSmod Subject (EDSS: 4.0±0.7; T25FW: 5.2±0.4 [s])

Fig. 3: **Panel plot illustrating the longitudinal severity estimate outcome for correctly classified HC, PwMSmild and PwMSmod subjects**. Depicted are the estimated level of disability for an example **(a)** HC subject; **(b)** a PwMSmild subject; **(c)** a PwMSmod subject during the study. Each circle represents the severity outcome estimate for a 2MWT performed on a given date. Shaded blue lines depict the and 7-day trend, represented by the d-point centred moving average across days (d). Missing test dates (which are not depicted) were imputed using piecewise linear interpolation. Dashed black lines represent site-visits where the participant was assessed clinically.

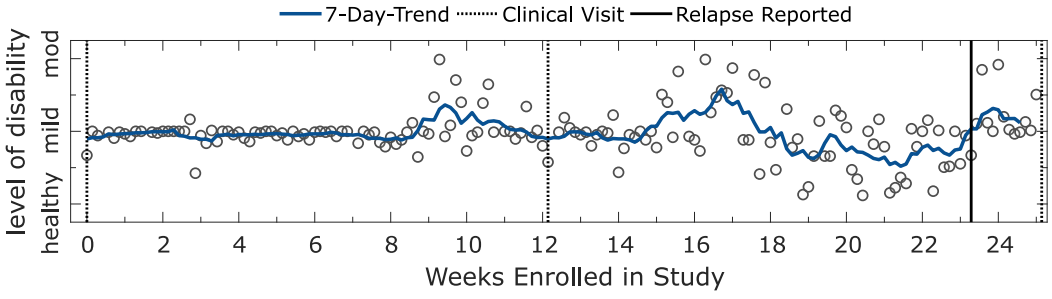
week 8 and week 12 is marked in figure 4d beginning with a long-dashed line. This PwMSmod subject was adherent to completing daily 2MWTs, with severity outcomes estimates consistently evaluated as moderately disabled up until week 8. Thereafter, the comparative number of completed daily 2MWTs dropped dramatically until study completion. It was observed that the stability of severity estimates predicted as PwMSmod diminished, with both greater variability between severity estimates and to the adherence of the participant to complete daily 2MWTs. Furthermore, a self-reported relapse was reported by this subject during week 22 using the FL application on their assigned smartphone, as marked by the solid black line.

IV. Discussion

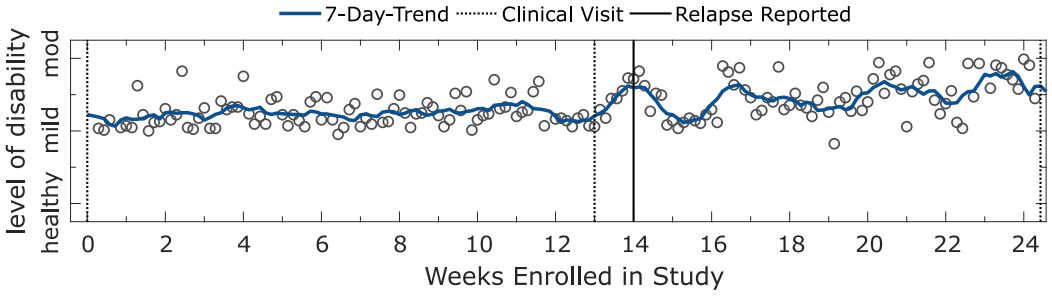
The FL PoC study demonstrates the capability of smartphone-based inertial sensor measurements to monitor ambulatory-related impairments during a remotely administered 2MWT to PwMS daily over a 24 week period. In this work, it was shown how a deep network classification model could (naïvely) estimate the level of participant disability from ordinal classification categories. Severity outcome estimates stratified across HC and PwMS groups and were strongly correlated to disease status ($r : 0.75$; $\rho : 0.71$, $p < 0.001$), as measured by the EDSS – considered the ground-truth assessment in PwMS [25]. For instance, no misclassifcation of HC as PwMSmod was observed, or vice-versa, indicating that severity estimates were reflective of true disease status (figure 2). More interestingly, those subjects at classification boundaries displayed severities representative of their clinical assessments. For instance, those with EDSS just above 3.5 (i.e. PwMSmod) were misclassified more as PwMSmild compared to those with EDSS much greater than 3.5, implying that a reflective estimate of disease severity could be captured by transforming a DCNN model into a simple probabilistic outcome per subject.



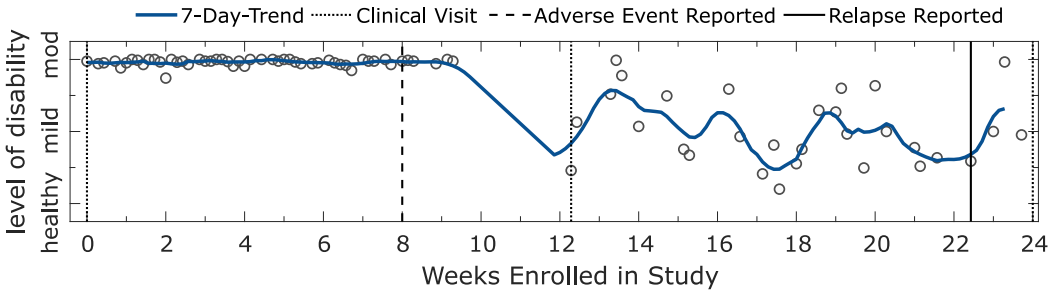
(a) Relapsing PwMSmild Subject (EDSS: 1.9±0.5; T25FW: 4.6±3.0 [s])



(b) Relapsing PwMSmild Subject (EDSS: 1.8±0.6; T25FW: 5.6±0.6 [s])



(c) Relapsing PwMSmod Subject (EDSS: 3.5±0; T25FW: 7.3±2.7 [s])



(d) Relapsing PwMSmod Subject (EDSS: 3.8±0.3; T25FW: 9.9±2.0 [s])

Fig. 4: **Panel plot illustrating the longitudinal severity estimate outcomes for participants who self-reported a relapse using the FLOODLIGHT smartphone application during the study**. Each circle represents the severity outcome estimate for a 2MWT performed on a given date. Shaded blue lines depict the and 7-day trend, represented by the d-point centred moving average across days (d). Missing test dates (which are not depicted) were imputed using piecewise linear interpolation. Dashed black lines represent site-visits where the participant was assessed clinically. Dates of self-reported relapse onset are represented in black. Note: the participant in figure 4d also reported (non-relapse) adverse clinical events occurring on non-specified dates between weeks 8 and 12.

A. Examining Participant-level Longitudinal Trends

The longitudinal patterns of healthy controls versus participants with varying manifestations of MS severity could be characterised by examining severity outcomes over the duration of the FL study for individual subjects. For instance, figure 3 depicted examples of stable trends for correctly classified HC, PwMSmild and PwMSmod subjects respectively. While participants had some incorrect predictions, the mean severity prediction over all repeated tests reflected the participant's true class grouping.

Evaluating subject's performance longitudinally suggested that severity estimates may be sensitive to capture MS-symptom worsening. An intriguing observation related to the stable PwMSmod participant depicted in figure 3c, who was mainly predicted with a severity of PwMSmod, with a relatively

consistent 7-day average. Some sequences of tests were predicted as milder however, particularly before the midway clinical visit in week 13. Interestingly, after week 13, this subjects' EDSS rose by $+1$ to 4.5. A Brown–Forsythe (BF) test demonstrated that this subject had greater variance in their severity outcome before this clinical visit compared to after (BF, $p < 0.01$). Median severity outcomes were not significantly different between these time-points (Mann-Whitey U test, p=0.34), however mean severity outcomes were found to be significantly lower before this clinical visit than after (Welch's t-test: p<0.05). It should be noted however that a change in EDSS scores between clinical visits did not correspond to significant changes in ambulatory-based severity estimates for all participants.

B. Examining Participant-level Relapse Events

During the FL study, four participants experienced relapses which they self-reported using the application on their smartphones. Longitudinal analysis of the trajectories of daily severity estimates from these subjects revealed useful insights into the manifestation of relapses expressed in remote inertial sensor data. For instance, two subjects displayed an increased severity outcome up to and around the data of reporting a relapse (figure 4a and 4c), suggesting that sensor-based ambulatory outcomes could potentially be sensitive enough to remotely capture relapse events.

Observing the PwMSmild participant who reported a relapse (figure 4a), severity estimates increased after reporting a relapse, which corroborated with a worsening of clinically assessed symptoms from baseline (week 0; EDSS 1.5, T25FW: 4.9 [s]) to the unscheduled clinical visit, which was prompted by the relapse (EDSS: 2.5; T25FW, 7.5 [s]). Examination of severity outcomes leading up to week 3 demonstrated consistent "mild" trends using 7-day moving averages. Interestingly, after the date of onset of self-reported relapse, severity estimates rose towards "moderate", indicating that MS symptom manifestation had worsened. Longer term analysis demonstrated that there was a significantly higher variability in predicted severity outcomes after relapse date than before (BF, p< 0.001). This subject was further assessed during week 12, where their EDSS returned to as it was reported at baseline (EDSS 1.5; T25FW: N/A). Severity outcomes also returned to consistently "mild" towards the end of the study from weeks 18 onwards, where median (U test, p$= 0.24$) and mean (Welch's t-test, p$= 0.13$) severity outcomes where not significantly different before- and after-relapse. This subject was predicted as PwMSmild over their entire 2MWT outcome measures.

In contrast, the example participant presented in figure 4b did not exhibit any significant changes in severity estimates around the date of reporting a relapse in week 23. However, it could also be noted that this subject's EDSS scores rose by +1 between week 12 and 24, and their ambulatory estimated outcomes were more variable after week 12 (BF, p< 0.01).

Figure 4c depicted a relapsing PwMSmod subject, with severity estimates that were consistently evaluated as "mild", up until week 13, where this participant reported a relapse on-setting using the FL application on their smartphone. Severity outcomes then increased towards "moderate" during week 13 and peaked at week 14, around the suspected relapse date reported at the end of week 13. Thereafter, severity outcomes stabilised to "mild" before becoming more variable and "moderate" until the end of the study. Considering the relapse reporting date as a threshold, it was found that severity outcomes were significantly "milder" before relapse (where severity outcomes evaluated as PwMSmild) than after relapse on-setting (where severity outcomes evaluated as PwMSmod) when testing between mean (Welch's t-test, p< 0.001) and between median (U test, p< 0.001) severity outcomes. A BF test also signified that severity outcome variability was higher after relapse on-setting than before (p< 0.01). This subject was misclassified as PwMSmild using all available 2MWTs, but interestingly was narrowly labelled a PwMSmod and not a PwMSmild subject using their available EDSS scores (EDSS, 3.5 ± 0).

Finally, figure 4d describes the longitudinal severity outcomes for a PwMSmod participant who was consistently estimated as having moderate disability for the first 9 weeks of the study period. During the mid-way assessment at week 12, this participant recalled that non-MS related adverse clinical events had occurred at unspecified points in the previous four weeks. Interestingly, adherence to executing daily 2MWTs dropped during this period, where a long-dashed line marks the beginning between weeks 8 and 12 in figure 4d. It was observed that after week 12, the variability in sensor-based ambulatory severity estimates increased, where predictions fluctuated between healthy and moderate. Furthermore, this participant was non-adherent at providing daily 2MWTs after week 12, in comparison than the first 9 weeks. Towards the end of the study, this participant then self-reported an MS-related relapse as having occurred in week 22. As such, we need to consider not only that sensor-based outcomes could remotely evaluate a patient's level of disability, but that an absence of available data itself might also be indicative of changes in disability status.

C. Limitations

Despite the potential of smartphone-based outcomes to remotely monitor individual participant's ambulatory function longitudinally, there are several limitations of this study which must be considered. Importantly, the severity outcomes explored in this work were naïve estimates; although outcomes captured a trend of increased impairment with higher severity (as modelled by EDSS, figure 2), they should not be considered an exact measure of MS, nor a surrogate clinical outcome to permit any clinical diagnosis, or replace in-clinic assessments.

It should also be noted that the estimated level of participant disability were not always accurate, there were many subject misclassifications, as evident in figure 2. Particularly, some HC were incorrectly estimated as MS, as well as some PwMSmod who were underestimated to have milder level of disability. In this work, we have only shown correct and stable estimate examples (figure 3), however, it must be noted that some participants, both healthy or with MS, followed irregular trends or whose estimated level of disability were consistently incorrect.

Planned future work will aim to further characterise misclassifications and participant variance. Given MS is heterogeneous disease, where symptoms fluctuate day-to-day, it must be considered that sometimes MS symptoms can be absent for a given day, or sequence of days. For instance, this may help explain why some PwMS participant 2MWTs can be evaluated healthy. It also must be acknowledged that severity estimates were based solely on 2MWT performance, an assessment originally only intended to investigate ambulatory function and fatigue in PwMS through the measurement of distance travelled [41]–[43]. Many participants in the FL PoC study may not have had ambulatory-related dysfunction, or whose milder level of disease did not impair their gait, compared to the healthy control cohort. As previously outlined, by definition PwMS with EDSS$<$3 may have little to no gait impairment [12], [15]. Furthermore, the blunt demarcation of mild and moderate MS based exclusively on the clinical EDSS score – which incorporates, but is not a direct measure of ambulatory function – could lead to an unreliable assignment of those "mild" versus "moderate" MS ambulatory function. For example, some participants might exhibit "moderate" symptoms that are more apparent in other functional domains, or have subtle alteration in ambulatory ability that a remote 2MWT assessment will not be sensitive to.

There are also several limitations associated with remote 2MWTs, which have been discussed previously in [12], [15], and must also be considered in the context of remotely estimating MS ambulatory severity. For example, although the 2MWT was standardised and analogous to that of an in-clinic performed assessment, the FL 2MWT was a remotely executed out-of-clinic assessment. As such, the performance of 2MWT can be highly influenced by the testing environment, such as the length of the hallways, the number and frequency of subject turns, or other factors which we cannot determine remotely [15].

In this work we proposed that averaging over categorical class predictions can create a simple and naïve estimate of ambulatory severity, but there could potentially be more informative and robust methodological approaches to learning disease severity estimates [44], [45]. It should be acknowledged that our DCNN model did not truly utilise the time-series nature of repeated 2MWT measurements from the FL PoC study. Each repeated test was treated as independent, and as such, trajectories did not incorporate any temporal information across a population or within a subject (for example, whether the previous day's $d-1$ test could affect the outcome at d or $d+1$). It would be assumed that this is critically missing temporal information which could help build more reliable and accurate longitudinal models, and should be considered as a key next step for future work. For instance, the repeated FL assessments, and therefore sensor outcomes that were extracted, could be analysed with models that exploit this aggregation of temporal information directly [46], [47]. Another limitation of averaging posterior class predictions is that we also average over uncertain or marginal predictions, often introducing a noise and variability into the unified estimate. Indeed, constructing more robust severity outcomes would not only explore more accurate modelling techniques, but should also aim to incorporate the data captured from other functional domains in FL, such as dexterity and cognition.

Nonetheless, we believe that the work presented in this study to be of important value, emphasising the potential of remote sensor outcomes to augment current in-clinic acquired patient information. The long-term remote monitoring of PwMS function could open up the space for true personalisation: the clustering of disease trajectories or similar patients, estimating the likelihood of disease progression, quantifying response to different treatments as a population or an individual, as well catching the mutable patterns of MS disease that are only visible out-of-clinic and as a function of time.

V. Conclusion

This work demonstrates the capability of smartphone technologies to administer daily ambulatory assessments to patients at home, and how that sensor data recorded can be transformed through state-of-the-art deep networks, to remotely monitor ambulatory-related level of disability over a 24 week period. The rapid development of frequent, objective, and sensitive digital measures of MS disability that can be administered remotely could revolutionise routine in-clinic assessments for PwMS. In the years to come, smartphone-based outcomes may identify and monitor digital signs of MS-related degeneration, ultimately informing better disease management techniques, to learn how different patients respond to various treatments, and potentially enabling the development of personalised therapeutic interventions.

Appendix A
Participant 2MWT adherence

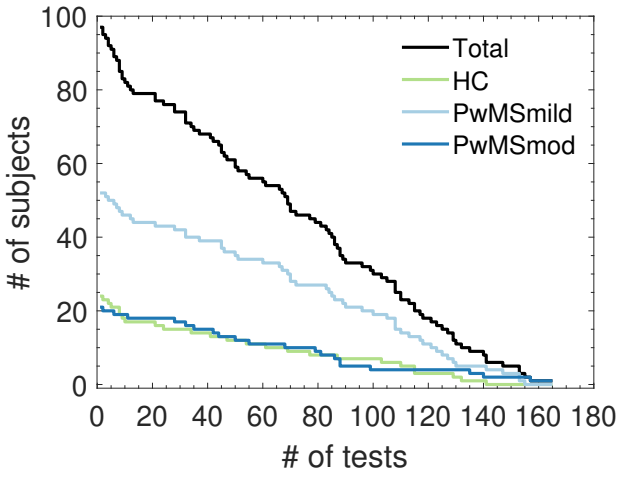


Fig. A.1: **Participant adherence rates**. Each line depicts the number valid Two Minute Walk Test (2MWT) recordings contributed for each subject group for the study duration.

Acknowledgements

The authors would like to thank all staff and participants involved in capturing test data. This study was sponsored by F. Hoffmann-La Roche Ltd. This research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). This research also

received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. During the completion of this work, A. P. Creagh was a Ph.D. student at the University of Oxford and acknowledges the support of F. Hoffmann-La Roche Ltd.; F. Dondelinger and F. Lipsmeier are employees of F. Hoffmann-La Roche Ltd; M. Lindemann is a consultant for F. Hoffmann-La Roche Ltd. via Inovigate; M. De Vos has nothing to disclose.

REFERENCES

[1] G. Comi, M. Filippi, F. Barkhof, *et al.*, "Effect of early interferon treatment on conversion to definite multiple sclerosis: A randomised study," *The Lancet*, vol. 357, no. 9268, pp. 1576–1582, 2001.

[2] S. L. Hauser and D. S. Goodin, "Multiple sclerosis and other demyelinating diseases," in *Harrison's Principles of Internal Medicine, 19e*, D. Kasper, A. Fauci, S. Hauser, *et al.*, Eds. New York, NY: McGraw-Hill Education, 2014.

[3] F. D. Lublin and S. C. Reingold, "Defining the clinical course of multiple sclerosis: Results of an international survey," *Neurology*, vol. 46, no. 4, pp. 907–911, 1996.

[4] M. M. Goldenberg, "Multiple sclerosis review," *Pharmacy and Therapeutics*, vol. 37, no. 3, p. 175, 2012.

[5] C. Confavreux, S. Vukusic, T. Moreau, and P. Adeleine, "Relapses and progression of disability in multiple sclerosis," *New England Journal of Medicine*, vol. 343, no. 20, pp. 1430–1438, 2000.

[6] E. Fisher, J.-C. Lee, K. Nakamura, and R. A. Rudick, "Gray matter atrophy in multiple sclerosis: A longitudinal study," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 64, no. 3, pp. 255–265, 2008.

[7] L. Steinman, "Multiple sclerosis: A two-stage disease," *Nature immunology*, vol. 2, no. 9, pp. 762–764, 2001.

[8] A. P. Creagh, C. Simillion, A. Scotland, *et al.*, "Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the draw a shape test," *Physiological Measurement*, vol. 41, no. 5, p. 054 002, 2020.

[9] A. Pratap, D. Grant, A. Vegesna, *et al.*, "Evaluating the utility of smartphone-based sensor assessments in persons with multiple sclerosis in the real-world using an app (elevatems): Observational, prospective pilot digital health study," *JMIR mHealth and uHealth*, vol. 8, no. 10, e22108, 2020.

[10] L. Pham, T. Harris, M. Varosanec, *et al.*, "Smartphone-based symbol-digit modalities test reliably captures brain damage in multiple sclerosis," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–13, 2021.

[11] X. Montalban, J. Graves, L. Midaglia, *et al.*, "A smartphone sensor-based digital outcome assessment of multiple sclerosis," *Multiple Sclerosis Journal*, p. 13 524 585 211 028 561, 2021.

[12] A. P. Creagh, F. Lipsmeier, M. Lindemann, and M. De Vos, "Interpretable deep learning for the remote characterisation of ambulation in multiple sclerosis using smartphones," *arXiv preprint arXiv:2103.09171*, 2021.

[13] R. Bove, C. C. White, G. Giovannoni, *et al.*, "Evaluating more naturalistic outcome measures: A 1-year smartphone study in multiple sclerosis," *Neurol Neuroimmunol Neuroinflamm*, vol. 2, no. 6, e162, 2015.

[14] G. Brichetto, L. Pedullà, J. Podda, and A. Tacchino, "Beyond center-based testing: Understanding and improving functioning with wearable technology in ms," *Multiple Sclerosis Journal*, vol. 25, no. 10, pp. 1402–1411, 2019.

[15] A. P. Creagh, C. Simillion, A. Bourke, *et al.*, "Smartphone- and smartwatch-based remote characterisation of ambulation in multiple sclerosis during the two-minute walk test," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[16] L. Midaglia, P. Mulero, X. Montalban, *et al.*, "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*, vol. 21, no. 8, e14863, 2019.

[17] J. J. Sosnoff, B. M. Sandroff, and R. W. Motl, "Quantifying gait abnormalities in persons with multiple sclerosis with minimal disability," *Gait & posture*, vol. 36, no. 1, pp. 154–156, 2012.

[18] C. L. Martin, B. Phillips, T. Kilpatrick, *et al.*, "Gait and balance impairment in early multiple sclerosis in the absence of clinical disability," *Multiple Sclerosis Journal*, vol. 12, no. 5, pp. 620–628, 2006.

[19] S. Crenshaw, T. Royer, J. Richards, and D. Hudson, "Gait variability in people with multiple sclerosis," *Multiple Sclerosis Journal*, vol. 12, no. 5, pp. 613–619, 2006.

[20] J. M. Huisinga, M. Mancini, R. J. S. George, and F. B. Horak, "Accelerometry reveals differences in gait variability between patients with multiple sclerosis and healthy controls," *Annals of biomedical engineering*, vol. 41, no. 8, pp. 1670–1679, 2013.

[21] R. I. Spain, M. Mancini, F. B. Horak, and D. Bourdette, "Body-worn sensors capture variability, but not decline, of gait and balance measures in multiple sclerosis over 18 months," *Gait & posture*, vol. 39, no. 3, pp. 958–964, 2014.

[22] R. W. Motl, B. M. Sandroff, Y. Suh, and J. J. Sosnoff, "Energy cost of walking and its association with gait parameters, daily activity, and fatigue in persons with mild multiple sclerosis," *Neurorehabilitation and neural repair*, vol. 26, no. 8, pp. 1015–1021, 2012.

[23] H. L. Zwibel, "Contribution of impaired mobility and general symptoms to the burden of multiple sclerosis," *Advances in therapy*, vol. 26, no. 12, pp. 1043–1057, 2009.

[24] L. Hemmett, J. Holmes, M. Barnes, and N. Russell, "What drives quality of life in multiple sclerosis?" *Qjm*, vol. 97, no. 10, pp. 671–676, 2004.

[25] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis an expanded disability status scale (edss)," *Neurology*, vol. 33, no. 11, pp. 1444–1444, 1983.

[26] R. Rudick, G. Cutter, and S. Reingold, "The multiple sclerosis functional composite: A new clinical outcome

measure for multiple sclerosis trials," *Multiple Sclerosis Journal*, vol. 8, no. 5, pp. 359–365, 2002.

[27] R. W. Motl, J. A. Cohen, R. Benedict, *et al.*, "Validity of the timed 25-foot walk as an ambulatory performance outcome measure for multiple sclerosis," *Mult Scler*, vol. 23, no. 5, pp. 704–710, 2017.

[28] R. W. Motl, Y. Suh, S. Balantrapu, *et al.*, "Evidence for the different physiological significance of the 6-and 2-minute walk tests in multiple sclerosis," *BMC neurology*, vol. 12, no. 1, p. 6, 2012.

[29] M. Sparaco, L. Lavorgna, R. Conforti, *et al.*, "The role of wearable devices in multiple sclerosis," *Multiple sclerosis international*, vol. 2018, 2018.

[30] D. Jarchi, J. Pope, T. K. Lee, *et al.*, "A review on accelerometry based gait analysis and emerging clinical applications," *IEEE Reviews in Biomedical Engineering*, 2018.

[31] A. Godfrey, R. Conway, D. Meagher, and G. Ó'Laighin, "Direct measurement of human movement by accelerometry," *Medical engineering & physics*, vol. 30, no. 10, pp. 1364–1386, 2008.

[32] B. R. Greene, S. Rutledge, I. McGurgan, *et al.*, "Assessment and classification of early-stage multiple sclerosis with inertial sensors: Comparison against clinical measures of disease state," *IEEE journal of biomedical and health informatics*, vol. 19, no. 4, pp. 1356–1361, 2015.

[33] R. Spain, R. S. George, A. Salarian, *et al.*, "Body-worn motion sensors detect balance and gait deficits in people with multiple sclerosis who have normal walking speed," *Gait & posture*, vol. 35, no. 4, pp. 573–578, 2012.

[34] M. Psarakis, D. A. Greene, M. H. Cole, *et al.*, "Wearable technology reveals gait compensations, unstable walking patterns and fatigue in people with multiple sclerosis," *Physiological measurement*, vol. 39, no. 7, p. 075 004, 2018.

[35] A. K. Bourke, A. Scotland, F. Lipsmeier, *et al.*, "Gait characteristics harvested during a smartphone-based self-administered 2-minute walk test in people with multiple sclerosis: Test-retest reliability and minimum detectable change," *Sensors*, vol. 20, no. 20, 2020.

[36] T. T. Um, F. M. Pfister, D. Pichler, *et al.*, "Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks," in *Proceedings of the 19th ACM international conference on multimodal interaction*, 2017, pp. 216–220.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[39] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[40] M. B. Brown and A. B. Forsythe, "Robust tests for the equality of variances," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 364–367, 1974.

[41] D. A. Scalzitti, K. J. Harwood, J. R. Maring, *et al.*, "Validation of the 2-minute walk test with the 6-minute walk test and other functional measures in persons with multiple sclerosis," *International journal of MS care*, vol. 20, no. 4, pp. 158–163, 2018.

[42] B. C. Kieseier and C. Pozzilli, "Assessing walking disability in multiple sclerosis," *Multiple Sclerosis Journal*, vol. 18, no. 7, pp. 914–924, 2012, PMID: 22740603.

[43] D. Gijbels, B. Eijnde, and P. Feys, "Comparison of the 2- and 6-minute walk test in multiple sclerosis," *Multiple Sclerosis Journal*, vol. 17, no. 10, pp. 1269–1272, 2011, PMID: 21642370.

[44] K. Dyagilev and S. Saria, "Learning (predictive) risk scores in the presence of censoring due to interventions," *Machine Learning*, vol. 102, no. 3, pp. 323–348, 2016.

[45] A. Zhan, S. Mohan, C. Tarolli, *et al.*, "Using smartphones and machine learning to quantify parkinson disease severity: The mobile parkinson disease score," *JAMA neurology*, vol. 75, no. 7, pp. 876–880, 2018.

[46] P. Schwab and W. Karlen, "Phonemd: Learning to diagnose parkinson's disease from smartphone data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1118–1125.

[47] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.