*Original Research Paper*

# Real-world keystroke dynamics are a potentially valid biomarker for clinical disability in multiple sclerosis

**KH Lam** [iD], **KA Meijer, FC Loonstra** [iD], **EME Coerver** [iD], **J Twose** [iD], **E Redeman,**
**B Moraal, F Barkhof** [iD], **V de Groot, BMJ Uitdehaag and J Killestein**

## Abstract

**Background:** Clinical measures in multiple sclerosis (MS) face limitations that may be overcome by utilising smartphone keyboard interactions acquired continuously and remotely during regular typing.
**Objective:** The aim of this study was to determine the reliability and validity of keystroke dynamics to assess clinical aspects of MS.
**Methods:** In total, 102 MS patients and 24 controls were included in this observational study. Keyboard interactions were obtained with the Neurokeys keyboard app. Eight timing-related keystroke features were assessed for reliability with intraclass correlation coefficients (ICCs); construct validity by analysing group differences (in fatigue, gadolinium-enhancing lesions on magnetic resonance imaging (MRI), and patients vs controls); and concurrent validity by correlating with disability measures.
**Results:** Reliability was moderate in two (ICC=0.601 and 0.742) and good to excellent in the remaining six features (ICC=0.760–0.965). Patients had significantly higher keystroke latencies than controls. Latency between key presses correlated the highest with Expanded Disability Status Scale ($r=0.407$) and latency between key releases with Nine-Hole Peg Test and Symbol Digit Modalities Test ($\rho=0.503$ and $r=-0.553$, respectively), $ps<0.001$.
**Conclusion:** Keystroke dynamics were reliable, distinguished patients and controls, and were associated with clinical disability measures. Consequently, keystroke dynamics are a promising valid surrogate marker for clinical disability in MS.

*Keywords:* Multiple sclerosis, disability evaluation, upper extremity, cognition, smartphone, touch typing, ambulatory monitoring, ecological momentary assessment

Correspondence to:
**KH Lam**
Department of Neurology,
Amsterdam University
Medical Centers, Vrije
Universiteit Amsterdam,
Amsterdam Neuroscience,
MS Center Amsterdam,
De Boelelaan 1117, 1081
HV Amsterdam, The
Netherlands.
**k.lam1@amsterdamumc.nl**

**KH Lam**
**FC Loonstra**
**EME Coerver**
**BMJ Uitdehaag**
**J Killestein**
Department of Neurology,
Amsterdam University
Medical Centers, Vrije
Universiteit Amsterdam,
Amsterdam Neuroscience,
MS Center Amsterdam,
Amsterdam, The Netherlands

**KA Meijer**
**J Twose**
**E Redeman**
Neurocast B.V., Amsterdam,
The Netherlands

**B Moraal**
Department of Radiology
and Nuclear Medicine,
Amsterdam University
Medical Centers, Vrije
Universiteit Amsterdam,
Amsterdam Neuroscience,
MS Center Amsterdam,
Amsterdam, The Netherlands

**F Barkhof**
Department of Radiology
and Nuclear Medicine,
Amsterdam University
Medical Centers, Vrije
Universiteit Amsterdam,
Amsterdam Neuroscience,
MS Center Amsterdam,
Amsterdam, The
Netherlands/Queen Square
Institute of Neurology and
Centre for Medical Image
Computing, University
College London, London, UK

**V de Groot**
Department of Rehabilitation
Medicine, Amsterdam
University Medical
Centers, Vrije Universiteit
Amsterdam, MS Center
Amsterdam, Amsterdam
Neuroscience, Amsterdam,
The Netherlands

## Introduction

Multiple sclerosis (MS) is a heterogeneous immune-mediated disease characterised by inflammatory disease activity and progression of disability over time.[1] Among the many clinical manifestations, fatigue is one of the most common symptoms in MS with its severity often negatively impacting quality of life, work and social participation.[2] Clinical deficits underlying ongoing inflammatory activity can be subtle or even absent, while responsive and standardised assessment tools for disability worsening and fatigue are lacking.[3,4] Hence, the assessment of disease activity, disability and fatigue relies heavily on magnetic resonance imaging (MRI), Expanded Disability Status Scale (EDSS) and multiple patient-reported fatigue outcomes, respectively.[5] However,

the current clinical measures are limited in assessment frequency, are obtrusive, are momentary and in the case of the EDSS are faced with inter- and intra-rater variability.[5]

Alternatively, remote health monitoring through devices with sensors allows continuous, objective and potentially more time- and cost-efficient assessment.[6] Of such devices, smartphones are increasingly used in healthcare and widely available;[7] 87% of patients with MS aged 20–72 years own a smartphone.[8] Typing is a common user–device interaction requiring coordinated, repetitive motor skills (e.g. coordination and manual dexterity) and higher order cognitive functions (e.g. information processing and attention), both of which tend to be affected in MS.[9] Quantitative

analysis of press and release keyboard interactions – *keystroke dynamics* – allows the identification of typing behaviour (e.g. amount, speed and error rate of typing) in a real-world setting which can reflect motor and non-motor functioning.[10,11] We, therefore, hypothesise that keystroke dynamics can be utilised as a biomarker in MS in relation to disease activity, clinical disability and fatigue.

Our objective is to investigate (1) the feasibility and reliability of real-world keystroke dynamics collected by smartphone technology and (2) its validity to assess fatigue, MRI disease activity and clinical disability in MS.

## Methods

### Study design
This is a prospective observational cohort study at Amsterdam University Medical Centers, location VU University Medical Center. The study comprises five clinical visits with 3-month intervals during which keyboard interaction data were remotely obtained from the everyday environment of the participants. Reported here are the results of the pre-planned analysis of the first clinical visit ($M_0$), which included patient-reported outcomes 2 weeks later ($M_{0+2wk}$). Participants with missing clinical assessments or less than 11 days (i.e. <75%) of keystroke data between $M_0$ and $M_{0+2wk}$ were excluded from the analysis. Study approval was granted by the local institutional ethics review board and the institutional data protection officer conforming to the European General Data Protection Regulation. In compliance with Dutch legislation regarding clinical research involving medical devices, the Dutch healthcare inspectorate has been notified of the study. Written informed consent was obtained from all participants.

### Participants
Patients with MS and healthy controls (HC) were included from August 2018 to December 2019. Inclusion criteria were regular use of a smartphone with Android or iOS, age between 18 and 65 years and, for the patient group, a definite diagnosis of MS.[12] Exclusion criteria were EDSS score of 7.5 or higher, clinical disease activity or changes in disease-modifying drugs in the past 2 months, significant visual or upper extremity deficit affecting the ability to type on a smartphone, and clinically significant mood, sleep or behavioural disorder based on medical history-taking by a screening physician.

### Main outcomes
Clinical measures spanning physical and mental domains were assessed at $M_0$: MRI, EDSS, Nine-Hole Peg Test (NHPT) and Symbol Digit Modalities Test (SDMT). MRI of the brain included T1-weighted images after administration of gadolinium (Gd). The MRI images were assessed by a neuroradiologist for the presence of Gd-enhancing demyelinating lesions, indicative of inflammatory disease activity. The EDSS is a measure of severity of disability due to MS and was performed by trained physicians.[13] The NHPT and SDMT were used as measures for manual dexterity and information processing speed, respectively.[14,15] At $M_{0+2wk}$, the Fatigue Severity Scale (FSS) and Checklist Individual Strength Fatigue subscale (CIS-F) were assessed. The FSS and CIS-F measures self-reported fatigue over the past 7 and 14 days, respectively, with higher scores corresponding with more severe fatigue.[16,17] The CIS-F cutoff of ≥35 was used to indicate severe fatigue.[18]

### Keystroke dynamics
A smartphone app (Neurokeys, Neurocast B.V., Amsterdam) was developed for Android and iOS to measure health status through regular typing on the smartphone. The Neurokeys keyboard was installed on the participants' personal smartphone and replaced the default keyboard. During regular typing, keyboard interactions of interest were logged and timestamped in the background: alphanumeric keys, backspaces, space bars and punctuation keys. Based on these timestamped key types, the manner and rhythm of typing can be discerned by analysing keystroke features. Keyboard interactions were continuously collected and stored per typing session, defined as one successive period of activation followed by inactivation of the keyboard. When a typing session starts, keystroke data from the previous typing session are sent and removed from the smartphone. The collection of keystroke data through Neurokeys did not require additional action from the participants.

### Definition and aggregation of keystroke features
Between $M_0$ and $M_{0+2wk}$, general typing characteristics were obtained: total number of interactions, typing session length, word length, number of backspaces and number of space bars. From the keyboard interactions, eight timing-related keystroke features were derived (Figure 1). Features based on alphanumeric keys were latency between successive key presses (*Press-Press Latency*) and releases (*Release-Release Latency*), time between a key press and subsequent
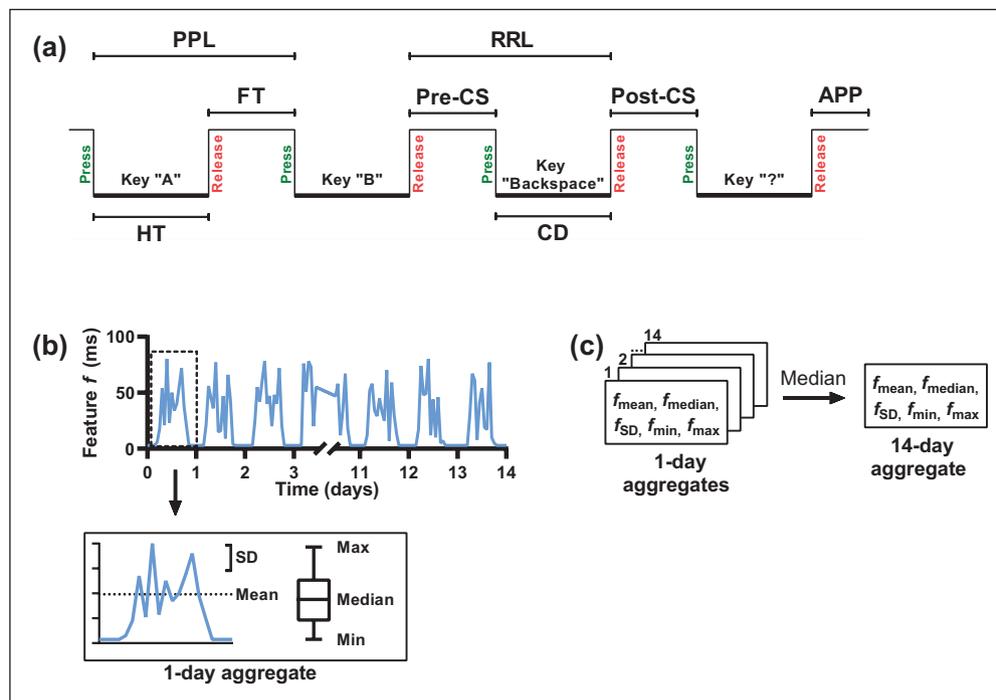
**Figure 1.** Schematic representation of the definition and aggregation of the timing-related keystroke features. (a) Keystroke features were derived from timestamped keyboard interactions. (b) For each keystroke feature, five summary statistics (i.e. vectors) were calculated to aggregate all typing sessions per day. (c) The five vectors for each feature were then aggregated into 14 days by taking the median value.
PPL: Press-Press Latency; RRL: Release-Release Latency; FT: Flight Time; Pre-CS: Pre-Correction Slowing; Post-CS: Post-Correction Slowing; APP: After Punctuation Pause; HT: Hold Time; CD: Correction Duration.

release (*Hold Time*) and time between a key release and the next key press (*Flight Time*). Features related to the backspace key were time before (*Pre-Correction Slowing*), during (*Correction Duration*) and after a backspace key press (*Post-Correction Slowing*). Feature based on punctuation marks was time between a punctuation mark and a subsequent alphanumeric key (*After Punctuation Pause*).

To match the keystroke features gathered per typing session with the clinical measures, aggregation into 14 days (and 7 days for comparison with the FSS) was performed (see Figure 1). Typing sessions were first aggregated per day. To aggregate the high sample rate data while retaining meaningful information, for each feature typing sessions were aggregated by calculating five summary statistics (i.e. vectors): mean and median (indicators of central tendency), standard deviation (SD; indicator of dispersion) and minimum and maximum (indicators of range). Then, the 14-day aggregate was obtained by taking the median value of the daily aggregates. Finally, to limit the number of tests, one composite score was computed for each 14-day aggregated keystroke feature by normalising the five vectors $v$ into $z$-scores with $z_v = (x_v - \mu_v) / \sigma_v$ and averaging

the five $z$-scores. In this formula, $z_v$ is the $z$-score, $x_v$ is a value of one participant, and $\mu_v$ and $\sigma_v$ are the mean and SD, respectively, of all participants.

### Statistical analysis

Keystroke features were assessed for normality using the Kolmogorov–Smirnov test. Non-normally distributed features were transformed using the 10th logarithm, natural logarithm or Box-Cox transformation.[19] Test–retest reliability of the keystroke features was assessed using 14-day keystroke aggregates prior to and after $M_{0+2wk}$, as no substantial change was expected between the two time windows. Intraclass correlation coefficients (ICCs) for consistency were calculated using a two-way random-effects model. Estimates between 0.5 and 0.75, 0.75 and 0.9, and 0.9 or higher were indicative for moderate, good and excellent reliability, respectively.[20] The standard error of measurement (SEM) was quantified by $SEM = SD_{pooled} \times \sqrt{1 - ICC}$ .[21] Agreement between the test and retest period was visually assessed by constructing Bland–Altman plots with the mean differences and the 95% limits of agreement (mean difference ± 1.96 SD).[22]

Construct validity of the composite keystroke features was analysed through group comparisons with independent *t*-tests: HC and patients with MS, patients with non-severe (MS-NF) and severe fatigue (MS-F), and patients with and without Gd-enhancing lesions.[23] Subsequently, composite keystroke features with significant group differences were further explored to determine which of the five vectors were different between the groups. Concurrent validity was assessed by calculating Pearson's and Spearman's correlation coefficients between the composite keystroke features and clinical measures.[23] The 'static' clinical measures, EDSS, NHPT and SDMT, were correlated with 14-day keystroke aggregates. The retrospective FSS (past 7 days) and CIS-F (past 2 weeks) assessed at $M_{0+2wk}$ were correlated with 7- and 14-day keystroke aggregates, respectively. As recall bias in self-report measures tends to emphasise recent or more severe events, sensitivity analyses were performed post hoc to explore whether the fatigue measures correlated better with recent (i.e. 2 days prior to $M_{0+2wk}$) and extreme (i.e. maximum values across the 7- and 14-day aggregates) keystroke events compared to the overall 7- and 14-day aggregates.[24] The group and correlation analyses were performed with permutations to increase robustness of the results.[25] Permutation-adjusted *p*-values are reported, and values $<0.05$ were considered statistically significant.

## Results

During the recruitment period, 276 people were interested in the study, of whom 132 were not screened after receiving full study information: 91 were unresponsive or declined participation without specification, 31 declined participation due to time or effort constraints, 7 were unwilling to undergo MRI and 3 opted to participate in an intervention study. The remaining 144 people were screened for eligibility. However, 18 were excluded: no conventional use of or typing on the smartphone ($n=6$), age above 65 years ($n=5$), no definite diagnosis of MS ($n=4$) and presence of clinically significant depressive and sleeping disorder, visual impairment or severe tremor of the upper limbs (each $n=1$).

A total of 24 HC and 102 patients with MS were included. One HC dropped out, and two HC and one patient did not have complete $M_{0+2wk}$ clinical measurements. Of the remaining participants, 18 (85.7%) HC and 85 (84.2%) patients had sufficient keystroke data as defined and were included in the analysis. There were no significant differences in demographical and clinical characteristics between participants included in the analysis and the participants excluded

due to insufficient keystroke data. During the 14 days of follow-up, 96 participants had 14 active days, 10 were active for 13 days, 4 were active for 12 days and 1 had 11 active days of data (see also Supplemental Figure). Demographical and clinical characteristics are summarised in Table 1. Patients with MS had a mean age of 46.4 years, 75.3% were female, 60.0% had relapsing-remitting MS and the median EDSS score was 3.5 (range 1.5–7.0). There were no significant differences observed in age, sex distribution and level of education between HC and patients. HC had a shorter mean typing session length compared to patients with MS (Table 2). The remaining general typing characteristics were not different between HC and patients.

### Reliability

In patients with MS, ICC estimates (and SEM in parentheses) of the keystroke features were as follows: *Correction Duration* = 0.601 (0.181), indicating moderate reliability; *Hold Time* = 0.742 (0.219), *After Punctuation Pause* = 0.760 (0.202), *Post-Correction Slowing* = 0.787 (0.239), *Press-Press Latency* = 0.830 (0.241) and *Pre-Correction Slowing* = 0.865 (0.218), all indicative of good reliability; *Flight Time* = 0.965 (0.098) and *Release-Release Latency* = 0.965 (0.099) demonstrated excellent reliability. The SEM for the keystroke features was small compared to the SD of the normalised features. Figure 2 shows the Bland–Altman plots for the eight composite keystroke features. The mean differences (systematic error) between the test and retest period were small and ranged from −0.043 to 0.016. Post-Correction Slowing had the largest limits of agreement (−0.7 to 0.7), in which the differences were slightly more scattered with larger mean values. The limits of agreement of all the other seven keystroke features were relatively small with the magnitude of differences evenly distributed over the whole range of mean values.

### HC versus patients with MS

The keystroke features *Press-Press Latency, Release-Release Latency, Flight Time, Pre-Correction Slowing, Post-Correction Slowing* and *After Punctuation Pause* were higher in patients with MS compared to HC (Figure 3(a)). Of these six composite features, analyses using the individual statistical vectors revealed that the mean values of all six features and the median values of *Press-Press Latency, Release-Release Latency, Pre-Correction Slowing* and *Post-Correction Slowing* were significantly higher in patients than in HC. For *Pre-Correction Slowing* also, the SD and maximum were higher in patients compared to HC. For *After*

**Table 1.** Baseline demographical and clinical characteristics.

|  | HC (*n*=18) | Patients with MS (*n*=85) | *p* value |
|---|---|---|---|
| Age, years, mean (SD) | 45.2 (13.5) | 46.4 (10.1) | 0.720[a] |
| Sex, *n* (%) |  |  | 0.146[b] |
|    Female | 10 (55.6) | 64 (75.3) |  |
|    Male | 8 (44.4) | 21 (24.7) |  |
| Level of education, *n* (%) |  |  | 0.601[b] |
|    Low | 0 (0.0) | 2 (2.4) |  |
|    Middle | 4 (22.2) | 28 (32.9) |  |
|    High | 14 (77.8) | 55 (64.7) |  |
| MS type, *n* (%) |  |  | n.a. |
|    Relapsing-remitting |  | 51 (60.0) |  |
|    Secondary progressive |  | 25 (29.4) |  |
|    Primary progressive |  | 9 (10.6) |  |
| Disease duration, years, median (IQR) |  |  | n.a. |
|    Since diagnosis |  | 5.7 (3.0–13.5) |  |
|    Since onset |  | 11.3 (5.1–17.7) |  |
| EDSS, median (IQR) |  | 3.5 (2.5–4.0) | n.a. |
| NHPT, median (IQR) |  | 21.0 (19.4–23.9) | n.a. |
| SDMT, mean (SD) |  | 54.5 (10.4) | n.a. |
| FSS, median (IQR) | 2.0 (1.7–2.3) | 4.4 (3.4–5.1) | <0.001[c] |
| CIS-F, median (IQR) | 17.5 (13.0–24.0) | 35.0 (27.0–42.0) | <0.001[c] |

HC: healthy controls; MS: multiple sclerosis; EDSS: Expanded Disability Status Scale; NHPT: Nine-Hole Peg Test; SDMT: Symbol Digit Modalities Test; FSS: Fatigue Severity Scale; CIS-F: Checklist Individual Strength Fatigue subscale; SD: standard deviation; IQR: interquartile range.
[a]Independent *t*-test.
[b]Fisher's Exact test.
[c]Mann–Whitney *U*-test.

**Table 2.** General typing characteristics and keystroke features.

|  | HC (*n*=18) | Patients with MS (*n*=85) | *p* value |
|---|---|---|---|
| Total typing events, *n* | 1373.1 (976.9) | 1653.6 (1515.1) | 0.253[a] |
| Average typing session duration, ms | 17,343.5 (6376.1) | 21,854.8 (7411.2) | 0.018[b] |
| Average word length, characters | 4.2 (0.5) | 4.2 (0.4) | 0.757[b] |
| Ratio of backspaces,[c] % | 10.4 (4.4) | 9.2 (4.6) | 0.305[b] |
| Ratio of space bars,[c] % | 13.0 (1.4) | 13.1 (1.9) | 0.752[b] |
| Press-Press Latency, ms | −0.388 (0.357) | 0.082 (0.676) | 0.005[a] |
| Release-Release Latency, ms | −0.349 (0.395) | 0.074 (0.688) | 0.013[a] |
| Hold Time, ms | −0.172 (0.663) | 0.036 (0.551) | 0.163[a] |
| Flight Time, ms | −0.360 (0.418) | 0.076 (0.693) | 0.012[a] |
| Pre-Correction Slowing, ms | −0.511 (0.651) | 0.108 (0.785) | 0.002[a] |
| Post-Correction Slowing,[d] ms | 0.370 (0.164) | 0.478 (0.147) | 0.005[a] |
| Correction Duration,[e] ms | 0.505 (0.187) | 0.572 (0.221) | 0.234[a] |
| After Punctuation Pause, ms | −0.452 (0.627) | 0.096 (0.642) | 0.001[a] |

HC: healthy controls; SD: standard deviation; MS: multiple sclerosis.
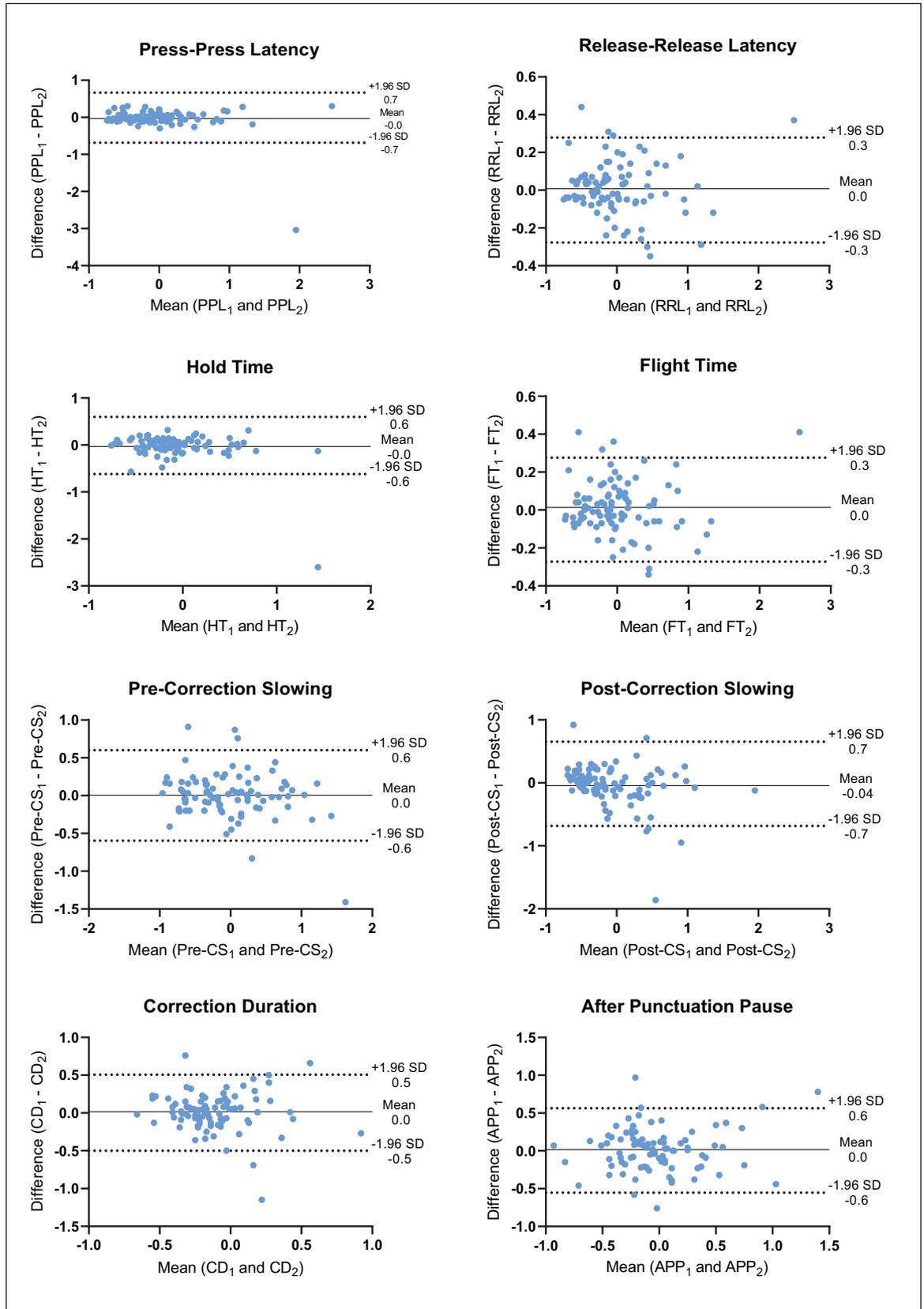All data are expressed as mean (SD).
[a]Independent *t*-tests.
[b]Mann–Whitney *U*-test.
[c]Ratio to total keystroke events.
[d]Box inverse transformed.
[e]Box square root transformed.

**Figure 2.** Bland–Altman plots of the keystroke features between the test–retest period. Between the test and retest period, the difference was plotted against the mean for each keystroke feature. The solid line represents the mean difference (systematic error) in keystroke features between the test–retest period, and the two dotted lines represent the 95% limits of agreement (random error).

*Punctuation Pause*, the SD was additionally higher in patients compared to HC. Thus, besides *Hold Time* and *Correction Duration*, timing-related keystroke features were significantly higher in patients compared to HC.

### Clinical disability

In patients with MS, EDSS was positively correlated with five of the eight timing-related keystroke features, of which latency between key presses showed the highest correlation, $r = 0.407$, $p < 0.001$ (Table 3, Figure 3(b)). NHPT was positively correlated with seven of the eight keystroke features, with the highest correlation observed with latency between key releases, $\rho = 0.503$, $p < 0.001$ (Table 3, Figure 3(c)). Finally, all eight keystroke features were negatively correlated with SDMT (Table 3). The highest correlation was found between *Release-Release Latency* and SDMT, $r = -0.553$, $p < 0.001$ (Figure 3(d)). Therefore, aside from the duration of backspaces for both NHPT and EDSS, and key hold duration and latency after punctuation marks for EDSS, all timing-related key press and release latencies were significantly associated with EDSS, NHPT and SDMT.

### Fatigue

In total, 39 (45.9%) patients with MS were stratified as 'non-severely fatigued' (MS-NF) and 46 (54.1%) as 'severely fatigued' (MS-F) using the CIS-F cutoff of 35. There were no differences in age, sex distribution, disease duration, NHPT and SDMT (all $p$s > 0.05) between MS-NF and MS-F. EDSS was lower in MS-NF, median (interquartile range (IQR)) = 3.0 (2.5–3.5), compared to MS-F, median (IQR) = 4.0 (3.0–4.5), $p = 0.020$. No differences in

timing-related keystroke features were observed between MS-NF and MS-F (all $p$s > 0.05). FSS and CIS-F scores did not correlate with the composite keystroke features (Table 3). Hence, the latency between key presses and releases was not different between patients with non-severe compared to severe fatigue, nor associated with level of fatigue. Post hoc sensitivity analyses only revealed significant correlations between FSS and *Hold Time* for recent ($r = 0.313$, $p = 0.005$) and severe ($r = 0.338$, $p = 0.001$) keystroke events.

### MRI Gd-enhancing lesions

Of the 51 patients with relapsing-remitting MS, 35 (68.6%) had no Gd-enhancing lesions, 12 (23.5%) had at least one Gd-enhancing lesion and 4 (7.8%) were not administered Gd (2 patients had a history of an allergic reaction to Gd, and in 2 patients, Gd administration was omitted). NHPT score was higher in the group with at least one Gd-enhancing lesion, median (IQR) = 21.7 (20.7–25.7), compared to the group without Gd-enhancing lesions, median (IQR) = 19.5 (18.3–21.6), $p = 0.003$. Besides NHPT, the two groups were similar in age, sex distribution, level of education, disease duration, EDSS and SDMT. There were no differences in timing-related keystroke features between patients without and with Gd-enhancing lesions (all $p$s > 0.05).

### Discussion

The results of this study show reliability and validity of keystroke dynamics to assess health status in MS. Generally, keystroke features collected over a 2-week period were different between patients with MS and HC. Group differences were driven by the mean and median values of latencies between key presses and

**Table 3.** Correlations coefficients between keystroke features and clinical measures in patients with MS ($n = 85$).

|  | PPL | RRL | HT | FT | Pre-CS | Post-CS | CD | APP |
|---|---|---|---|---|---|---|---|---|
| EDSS[a] | 0.407** | 0.380** | 0.150 | 0.383** | 0.300** | 0.352** | 0.207 | 0.209 |
| NHPT[b] | 0.455** | 0.503** | 0.251* | 0.457** | 0.386** | 0.441* | 0.179 | 0.408** |
| SDMT[a] | −0.525** | −0.553** | −0.286** | −0.525** | −0.300** | −0.444** | −0.164* | −0.317** |
| CIS-F[a] | −0.056 | −0.030 | 0.038 | −0.025 | −0.067 | −0.086 | −0.012 | −0.085 |
| FSS[a] | 0.041 | 0.080 | 0.030 | 0.028 | 0.006 | 0.066 | 0.038 | −0.216 |

PPL: Press-Press Latency; RRL: Release-Release Latency; HT: Hold Time; FT: Flight Time; Pre-CS: Pre-Correction Slowing; Post-CS: Post-Correction Slowing; CD: Correction Duration; APP: After Punctuation Pause; NHPT: Nine-Hole Peg Test; SDMT: Symbol Digit Modalities Test; FSS: Fatigue Severity Scale; CIS-F: Checklist Individual Strength Fatigue subscale; EDSS: Expanded Disability Status Scale.
[a]Pearson's correlation coefficient.
[b]Spearman's log-rank correlation coefficient.
*$p < 0.05$, **$p < 0.01$.
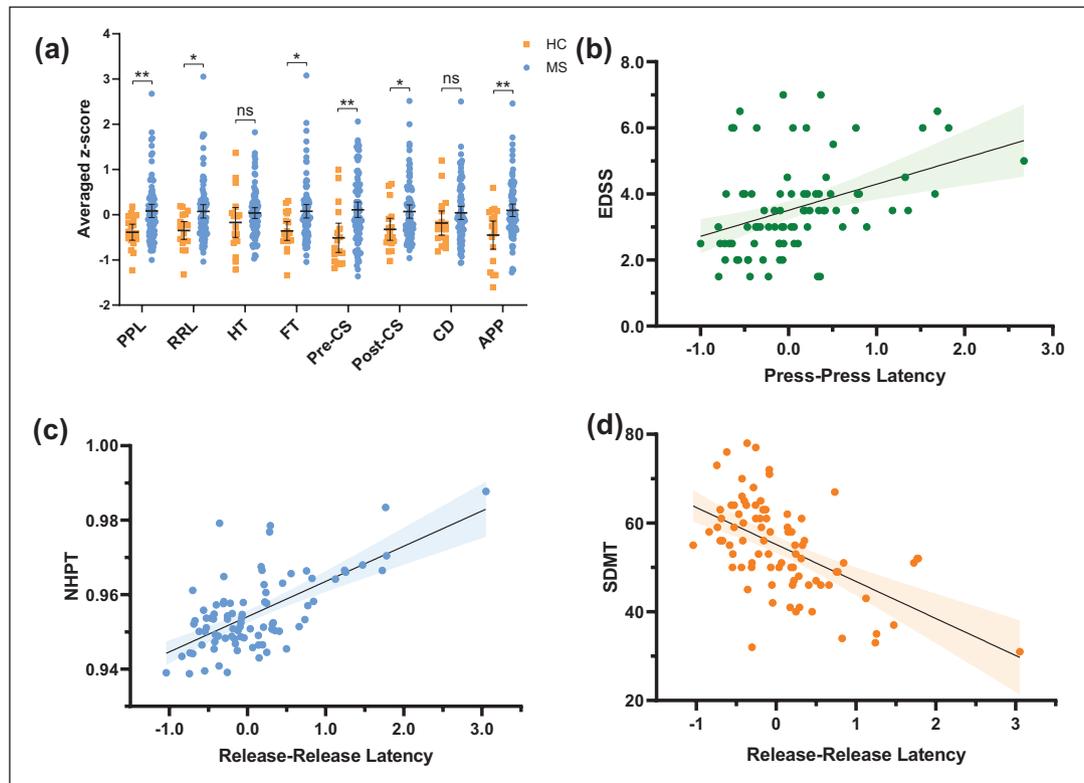
**Figure 3.** Scatter plots of timing-related keystroke features. (a) Keystroke features grouped between HC and patients with MS. Horizontal bars represent the mean, and error bars represent the 95% confidence interval of the mean. (b) Press-Press Latency (averaged z-score) and EDSS. (c) Release-Release Latency (averaged z-score) and NHPT (Box-Cox transformed). (d) Release-Release Latency (averaged z-score) and SDMT. HC: healthy controls; MS: patients with multiple sclerosis; PPL: Press-Press Latency; RRL: Release-Release Latency; HT: Hold Time; FT: Flight Time; Pre-CS: Pre-Correction Slowing; Post-CS: Post-Correction Slowing; CD: Correction Duration; APP: After Punctuation Pause; EDSS: Expanded Disability Status Scale; NHPT: Nine-Hole Peg Test; SDMT: Symbol Digit Modalities Test.
$^{ns}p \geqslant 0.05$, $^*p < 0.05$, $^{**}p < 0.01$.

releases, and the dispersion of latencies prior to backspaces and after punctuation marks, which were all higher in patients compared to HC. The keystroke features were also found to be associated with measures of clinical disability, information processing speed and manual dexterity. These findings together can be seen as a first step towards further clinical validation, showing that keystroke dynamics are a promising monitoring tool for disease status in MS.

Differences in keystroke features were found between patients and HC even with similarities in general typing characteristics and despite a relatively short disease duration (median of 5.7 years) and mild disease severity (median EDSS of 3.5) of our cohort. This suggests sensitivity of keystroke dynamics to subtle differences. Deficits in cognitive and motor skills are commonly seen and may already be present early in MS and can therefore explain the observed differences between patients and HC.[26] This is further supported by the correlations we found between keystroke

features and both NHPT and SDMT in our patients, in line with our hypothesis. For NHPT, our highest correlation was observed with the *Release-Release Latency*, in which the press and release finger motions resemble the grip and release finger movements during the NHPT. Similar results were reported with accelerometers during typing tasks on a physical keyboard and upper limb dysfunction.[27] The significant negative correlations found between the keystroke features and SDMT score, indicating longer keystroke delays, were associated with lower information processing speed, in line with a study comparing tactile touchscreen activity (swipe, tap and keystroke events) with neuropsychological constructs.[28] In addition to the NHPT and SDMT, a more general measure of clinical disability (the EDSS) and disease duration (data not shown) was also associated with timing-related keystroke features. Altogether, the moderate correlation coefficients between keystroke features and clinical disability measures suggest concurrent validity to some extent. This is intuitive given that

typing requires both cognitive and motor skills to perform coordinated, successive finger movements, producing intended words and sentences.[27,29]

We also hypothesised that fatigue negatively impacts typing. However, the keystroke features did not correlate with the FSS and CIS-F. Prior studies comparing momentarily assessed symptoms, including fatigue, with assessment of the symptom over a longer time period found none to modest correlations, possibly due to recall bias.[30,31] As recall bias was shown to be influenced by recent and extreme events, we performed post hoc sensitivity analyses by comparing recent and extreme keystroke data with the fatigue measures.[24] This, however, also did not yield consistent significant correlations. Changes in typing behaviour resulting from fatigue may be better described as fatigability, which is found to be a different construct than perceived fatigue.[32] In addition, and perhaps above all, our current cross-sectional analyses does not account for the between-subject variability, which is presumably high for a heterogeneous symptom such as fatigue. Hence, the use of keystroke dynamics to measure fatigue in MS needs further investigation in a longitudinal setting using more frequently assessed fatigue or fatigability measures. Similarly, in this cross-sectional analysis no significant differences were observed in keystroke features between patients with relapsing-remitting MS with and without Gd-enhancing lesions on MRI. Improvement in keystroke features may be anticipated in patients in whom disease activity diminishes. Thus, we expect to find within-subject differences over time in our longitudinal analyses for fatigue and for MRI Gd-enhancing lesions.

Smartphone keystroke dynamics have been investigated previously in Parkinson's disease, mood disorders and cognitive performance.[28,33,34] While these studies also showed associations between clinical measures and typing behaviour, most were conducted in an experimental setting by transcribing standardised text excerpts or using a specific type of smartphone. By contrast, keystroke data in our study were gathered during regular use of the participants' own smartphone. This ability to assess health outcomes more frequently and unobtrusively from a remote distance in real time is an important advantage compared to current clinical measures.[35] However, while interpreting our findings, limitations need to be considered. To comply with the assumptions of the statistical tests, data aggregation was performed, resulting in loss of detailed and high-frequency information regarding day-to-day differences. More complex models should be used next to utilise all available data points to determine the temporal association between keystroke features and clinical outcomes. Moreover, we focused on single keystroke features as a first step rather than a combination of keystroke features. Following this study, future directions should focus on overcoming these limitations, which were also among the challenges discussed recently on the use of wearable technology in MS.[36] As with all biomarkers, standardisation and reproducibility are required to further this (technological) biomarker into the clinical practice. Keystroke dynamics may then be employed in conjunction with the current clinical measures to supplement monitoring of MS outside of the clinical windows and more closely to the patients' functioning in the daily setting.

The results of this study show that keyboard interactions can be used as a monitoring tool through the collection of reliable high sample rate data. Based on the derived keystroke features, we were not only able to distinguish between the HC and MS group, but also show moderate correlations between the keystroke measures and clinical disability, manual dexterity and information processing speed in patients with MS. Therefore, keystroke dynamics are a promising potential outcome measure that enables remote and unobtrusive phenotyping of health status in patients with MS and opens up new applications for disease monitoring, patient management and outcomes for clinical trials.

### ORCID iDs

KH Lam https://orcid.org/0000-0003-0926-1445

FC Loonstra https://orcid.org/0000-0002-1644-4410

EME Coerver https://orcid.org/0000-0001-5673-9996

J Twose https://orcid.org/0000-0001-8499-0256

Frederik Barkhof https://orcid.org/0000-0003-3543-3706

### Supplemental material

Supplemental material for this article is available online.

### References

1. Compston A and Coles A. Multiple sclerosis. *Lancet* 2008; 372: 1502–1517.

2. Fisk JD, Pontefract A, Ritvo PG, et al. The impact of fatigue on patients with multiple sclerosis. *Canadian J Neurol Sci (Le Journal Canadien Des Sciences Neurologiques)* 1994; 21: 9–14.

3. Min M, Spelman T, Lugaresi A, et al. Silent lesions on MRI imaging – Shifting goal posts for treatment decisions in multiple sclerosis. *Mult Scler* 2018; 24(12): 1569–1577.

4. Uitdehaag BMJ. Disability outcome measures in phase III clinical trials in multiple sclerosis. *CNS Drugs* 2018; 32(6): 543–558.

5. Tur C, Moccia M, Barkhof F, et al. Assessing treatment outcomes in multiple sclerosis trials and in the clinical setting. *Nat Rev Neurol* 2018; 14(2): 75–93.

6. Majumder S, Mondal T and Deen MJ. Wearable sensors for remote health monitoring. *Sensors* 2017; 17: 130.

7. Mosa AS, Yoo I and Sheets L. A systematic review of healthcare applications for smartphones. *BMC Med Inform Decis Mak* 2012; 12: 67.

8. Griffin N and Kehoe M. A questionnaire study to explore the views of people with multiple sclerosis of using smartphone technology for health care purposes. *Disabil Rehabil* 2018; 40(12): 1434–1442.

9. Einarsson U, Gottberg K, von Koch L, et al. Cognitive and motor function in people with multiple sclerosis in Stockholm County. *Mult Scler* 2006; 12(3): 340–353.

10. Obaidat MS and Sadoun B. Verification of computer users using keystroke dynamics. *IEEE T Syst Man Cybern B Cybern* 1997; 27(2): 261–269.

11. Brizan DG, Goodkind A, Koch P, et al. Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. *Int J Human Comput Stud* 2015; 82: 57–68.

12. Thompson AJ, Banwell BL, Barkhof F, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol* 2018; 17: 162–173.

13. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). *Neurology* 1983; 33(11): 1444–1452.

14. Feys P, Lamers I, Francis G, et al. The nine-hole peg test as a manual dexterity performance measure for multiple sclerosis. *Mult Scler* 2017; 23(5): 711–720.

15. Benedict RH, DeLuca J, Phillips G, et al. Validity of the symbol digit modalities test as a cognition performance outcome measure for multiple sclerosis. *Mult Scler* 2017; 23(5): 721–733.

16. Krupp LB, LaRocca NG, Muir-Nash J, et al. The fatigue severity scale: Application to patients with multiple sclerosis and systemic lupus erythematosus. *Arch Neurol* 1989; 46(10): 1121–1123.

17. Vercoulen JH, Hommes OR, Swanink CM, et al. The measurement of fatigue in patients with multiple sclerosis: A multidimensional comparison with patients with chronic fatigue syndrome and healthy subjects. *Arch Neurol* 1996; 53(7): 642–649.

18. Worm-Smeitink M, Gielissen M, Bloot L, et al. The assessment of fatigue: Psychometric qualities and norms for the checklist individual strength. *J Psychosom Res* 2017; 98: 40–46.

19. Sakia RM. The box-cox transformation technique: A review. *J Royal Stat Soc Ser D* 1992; 41: 169–178.

20. Koo TK and Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropractic Med* 2016; 15: 155–163.

21. de Vet HCW, Terwee CB, Mokkink LB, et al. *Measurement in medicine: A practical guide*. Cambridge: Cambridge University Press, 2011.

22. Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1: 307–310.

23. Roach KE. Measurement of health outcomes: Reliability, validity and responsiveness. *J Prosthet Orthotic* 2006; 18: P8–P12.

24. Kahneman D, Fredrickson BL, Schreiber CA, et al. When more pain is preferred to less: Adding a better end. *Psychol Sci* 1993; 4: 401–405.

25. Romano JP and Wolf M. Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Stat Probab Lett* 2016; 113: 38–40.

26. Benedict RH, Holtzer R, Motl RW, et al. Upper and lower extremity motor function and cognitive impairment in multiple sclerosis. *J Int Neuropsychol Soc* 2011; 17(4): 643–653.

27. Londral A, Pinto S and de Carvalho M. Markers for upper limb dysfunction in amyotrophic lateral sclerosis using analysis of typing activity. *Clin Neurophysiol* 2016; 127(1): 925–931.

28. Dagum P. Digital biomarkers of cognitive function. *NPJ Digit Med* 2018; 1: 10.

29. Baus C, Strijkers K and Costa A. When does word frequency influence written production. *Front Psychol* 2013; 4: 963.

30. Heine M, van den Akker LE, Blikman L, et al. Real-time assessment of fatigue in patients with multiple sclerosis: How does it relate to commonly used self-report fatigue questionnaires. *Arch Phys Med Rehabil* 2016; 97(11): 1887–1894.

31. Stone AA, Schwartz JE, Broderick JE, et al. Variability of momentary pain predicts recall of weekly pain: A consequence of the peak (or salience) memory heuristic. *Pers Soc Psychol Bull* 2005; 31(10): 1340–1346.

32. Loy BD, Taylor RL, Fling BW, et al. Relationship between perceived fatigue and performance fatigability in people with multiple sclerosis: A systematic review and meta-analysis. *J Psychosom Res* 2017; 100: 1–7.

33. Iakovakis D, Hadjidimitriou S, Charisis V, et al. Touchscreen typing-pattern analysis for detecting fine motor skills decline in early-stage Parkinson's disease. *Sci Rep* 2018; 8: 7663.

34. Zulueta J, Piscitello A, Rasic M, et al. Predicting mood disturbance severity with mobile phone keystroke metadata: A biaffect digital phenotyping study. *J Med Intern Res* 2018; 20: e241.

35. Morris M, Intille SS and Beaudin JS. *Embedded assessment: Overcoming barriers to early detection with pervasive computing*. Berlin: Springer, 2005, pp. 333–346.

36. Brichetto G. We should monitor our patients with wearable technology instead of neurological examination: Commentary. *Mult Scler* 2020; 26(9): 1028–1030.

Visit SAGE journals online
journals.sagepub.com/
home/msj

Ⓢ SAGE journals