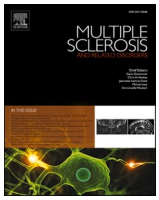
Multiple Sclerosis and Related Disorders 60 (2022) 103692

Contents lists available at ScienceDirect

Multiple Sclerosis and Related Disorders

journal homepage: www.elsevier.com/locate/msard



Original article

Towards individualized monitoring of cognition in multiple sclerosis in the digital era: A one-year cohort study

Ka-Hoo Lam [a,*], Ioan Gabriel Bucur [b], Pim Van Oirschot [c], Frank De Graaf [c], Hans Weda [c], Eva Strijbis [a], Bernard Uitdehaag [a], Tom Heskes [b], Joep Killestein [a], Vincent De Groot [d]

[a] Department of Neurology, Amsterdam University Medical Centers, Vrije Universiteit Amsterdam, MS Center Amsterdam, Amsterdam Neuroscience, Amsterdam, the Netherlands
[b] Institute for Computing and Information Sciences, Radboud University, Nijmegen, the Netherlands
[c] Orikami Digital Health Products, Nijmegen, the Netherlands
[d] Department of Rehabilitation Medicine, Amsterdam University Medical Centers, Vrije Universiteit Amsterdam, MS Center Amsterdam, Amsterdam Neuroscience, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Multiple sclerosis
Cognition
Outpatient monitoring
Smartphone
Digital technology
Patient-specific modeling

ABSTRACT

Background: Cognitive impairment is frequent in multiple sclerosis (MS), but reliable, sensitive and individualized monitoring in clinical practice is still limited. Smartphone-adapted tests may enhance the assessment of function as tests can be performed more frequently and within the daily living environment. The objectives were to prove reproducibility of a smartphone-based Symbol Digit Modalities Test (sSDMT), its responsiveness to relevant change in clinical cognitive outcomes, and develop an individual-based monitoring method for cognition.

Methods: In a one-year cohort study with 102 patients with MS, weekly sSDMTs were performed and analyzed on reproducibility parameters: the standard error of measurement (SEM) and smallest detectable change (SDC). Responsiveness of the sSDMT to relevant change in the 3-monthly clinically assessed SDMT (i.e. 4-point change) was quantified with the area under the receiver operating characteristic curve (AUC). Curve fitting of the weekly sSDMT scores of individual patients was performed with a local linear trend model to estimate and visualize the de-noised cognitive state and 95% confidence interval (CI). The optimal assessment frequency was determined by analyzing the CI bandwidth as a function of sSDMT assessment frequency.

Results: Weekly sSDMT showed improved reproducibility estimates (SEM=2.94, SDC=8.15) compared to the clinical SDMT. AUC-values did not exceed 0.70 in classifying relevant change in cSDMT. However, utilizing weekly sSDMT measurements, estimated state curves and the 95% CI were plotted showing detailed changes within individuals over time. With a test frequency of once per 12 days, 4-point changes in sSDMT can be detected.

Conclusion: A local linear trend model applied on sSDMT scores of individual patients increases the signal-to-noise ratio substantially, which improves the detection of statistically reliable changes. Therefore, this fine-grained individual-based monitoring approach can be used to complement current clinical assessment to enhance clinical care in MS.

Trial registration: Netherlands Trial Register NL7070; https://www.trialregister.nl/trial/7070

1. Introduction

Cognitive impairment is frequent in patients with multiple sclerosis (MS). (Benedict et al., 2020) The Symbol Digit Modalities Test (SDMT) is recommended for early screening and monitoring of cognitive function. (Kalb et al., 2018) Studies have shown reliability and validity for the SDMT in assessing information processing speed, one of the most commonly affected cognitive domains in MS. (Benedict et al., 2017) SDMT scores changes were also associated with change in employment status and relapses in MS. (SA Morrow et al., 2010; Benedict et al., 2021; Giedraitiene et al., 2018) However, individual-based monitoring of cognition is not yet routinely employed in clinical practice. This is due to

* Corresponding author at: Mailing address: De Boelelaan 1117 HV, Amsterdam, the Netherlands.
E-mail address: k.lam1@amsterdamumc.nl (K.-H. Lam).

https://doi.org/10.1016/j.msard.2022.103692
Received 14 December 2021; Received in revised form 4 February 2022; Accepted 18 February 2022
Available online 19 February 2022

lack of studies on the sensitivity to relevant change (i.e. responsiveness) on the individual level and relatively poor reproducibility of cognitive measures based on known studies. For instance, the smallest detectable change (SDC), a reproducibility parameter, of the SDMT exceeds 12 points, (Benedict et al., 2008; SA Morrow et al., 2010; Drake et al., 2010; Benedict et al., 2012; Sonder et al., 2014) a number that is much larger than the 4-point change found to be clinically meaningful. (Benedict et al., 2017) This indicates that even if clinically meaningful change occurs it is typically smaller in magnitude than, and therefore indistinguishable from, change due to extrinsic factors (e.g. mood, level of rest, rater or environmental factors). Therefore, more sensitive outcomes for monitoring cognition are needed.

Digital monitoring with the smartphone has several advantages over traditional clinical assessment. Smartphone assessment can be self-administered, reducing time spent by clinical personnel. Furthermore, measuring in a hospital environment is different from measuring in a real-life setting. Most importantly, smartphone measurements allow repeated assessments that could overall give a better indication of the patient's state over time compared to singular evaluations. Smartphone-based processing speed tests have recently been validated in a cross-sectional setting in MS. (Maillart et al., 2020; Pv et al., 2020; Pham et al., 2021; Lam et al., 2021) The SDC of this smartphone-adapted SDMT (sSDMT) was shown to be approximately 7 points, (Lam et al., 2021) which makes the sSDMT a promising tool for clinical application. However, studies on the sSDMT investigating its responsiveness in a longitudinal setting are lacking. Furthermore, methodological guidance on monitoring cognition using high frequency data in the clinical setting is absent, specifically on how to handle low signal-to-noise ratios or practice effects with repeated measurements, and how to interpret changes or determine the optimal assessment frequency.

1.1. Objectives

The objectives were to prove reproducibility of a smartphone-based SDMT, its responsiveness to relevant change in the clinical cognitive outcomes, and develop an individual-based monitoring method for cognition

2. Material and methods

The study comprised a single-center cohort study at Amsterdam University Medical Centers, location VU University Medical Center. (Netherlands Trial Register 2018) Patients with MS were recruited to use the MS sherpa® app to perform self-administered sSDMT assessments during a one-year follow-up. Clinical outcomes were assessed at three-monthly clinical visits: M_0 (baseline), M_3, M_6, M_9, and M_{12}. The interim analysis on validity and reliability of the sSDMT using the baseline data has been reported previously. (Lam et al., 2021) Patient characteristics were assessed at baseline: age, sex, level of education (low = primary school and/or low level secondary school, average = medium level secondary school, and high = high level secondary school and/or university degree), (Rijnen et al., 2020) MS type, disease duration, and disease severity quantified with the Expanded Disability Status Scale (EDSS) based on the neurological examination. (Kurtzke, 1983) Participants were consecutively included from August 2018 until a sample size of 100 patients was reached in December 2019. Eligibility criteria were age between 18 and 65 years, definite MS diagnosis, (Thompson et al., 2018) baseline EDSS score below 7.5, having an Android (5.0 or higher) or iOS (10 or higher) smartphone, no visual or upper extremity deficit affecting smartphone use, and no mood or sleep disorder impacting daily living assessed by a screening physician. The study received ethical approval (METc VUmc, 2017.576) and conformed to the Dutch legislation regarding data privacy and medical devices (VGR2006948). All patients gave written informed consent.

2.1. Clinical outcomes

During each clinical visit, the (oral) clinical SDMT (cSDMT) was assessed and alternated between two versions. (Benedict et al., 2012) At M_0 and M_{12}, the California Verbal Learning Test-II (CVLT-II, Dutch version) and the Brief Visuospatial Memory Test-Revised (BVMT-R) were assessed in addition to the cSDMT, as in the recommended Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS). (Langdon et al., 2012) For CVLT-II and BVMT-R, direct recall was assessed five and three times, respectively. Alternate versions were used for the CVLT-II and BVMT-R at M_0 and M_{12}. cSDMT score changes of $\geq$4 points were considered clinically relevant. (Benedict et al., 2017) For CVLT-II and BVMT-R no clinically meaningful change scores have been reported, a (distribution-based) cut-off of $\geq$0.5 SD of the overall sample at baseline was chosen instead. (Watt et al., 2021)

2.2. Measured smartphone SDMT

MS sherpa® (Orikami Digital Health Products, Nijmegen) is a system comprising a smartphone app (Android and iOS) for data collection, storage, and presentation, analysis algorithms, and clinician or researcher dashboard for user management and data visualization. (Orikami Digital Health Products 2021) Participants installed the MS sherpa® app on their own smartphone to self-perform the sSDMT. Using an on-screen numeric keypad, numbers corresponding to symbols were pressed according to nine symbol-digit combinations displayed at the top of the screen. Each trial had a 90-second duration and a randomized symbol-digit combination. The number of correct responses was scored as the measured sSDMT score. During the first four weeks the sSDMT was scheduled twice every three days. From week 5 onwards, the sSDMT was scheduled once every week. Push notifications were sent as reminders for scheduled tests.

2.3. Estimated smartphone SDMT

With frequent SDMT measurements, one gets confronted with natural variability in scores mostly due to day-to-day fluctuations rather than real cognitive change, as shown recently. (Pv et al., 2020; Pham et al., 2021) In order to obtain a better approximation of the true performance of cognition over time, a local linear trend model (LLTM, a type of linear-Gaussian state space model) was applied on the repeated sSDMT measurements for individual patients. (Durbin and Koopman, 2012) This data-driven curve fitting approach (patent pending: NL2028255) de-noises the measurements by taking advantage of high frequency data collection provided by digital monitoring. The local linear trend fit to the sSDMT time series, models variations in the level and slope, and irregular residual variation. The level and slope variations characterize the trend, a slowly varying component interpretable as real change in processing speed. The residual variation characterizes the error or disturbance in measurement. (Pv et al., 2020; Pham et al., 2021)

This state space model can be seen as an extension of (maximum likelihood) factor analysis. The measured sSDMT scores can be considered noisy measurements of the underlying latent state factor (i.e. processing speed), described as:

$$y_i^t = \mu_i^t + \varepsilon_i^t.$$

For each patient i at time t, y_i^t is the measured sSDMT score which is equal to the latent sSDMT state (μ_i^t) plus (zero-mean Gaussian) noise denoted by ε_i^t (see Fig. 1). The noise is assumed to be stationary, but having a different distribution for each patient. Time is indicated by the discrete index t, which is the number of days from inclusion. Days without measurements were treated as missing. The zero-mean Gaussian noise closely relates to measurement error and comprises short-term irregular effects due to factors such as time of the day, mood, level of

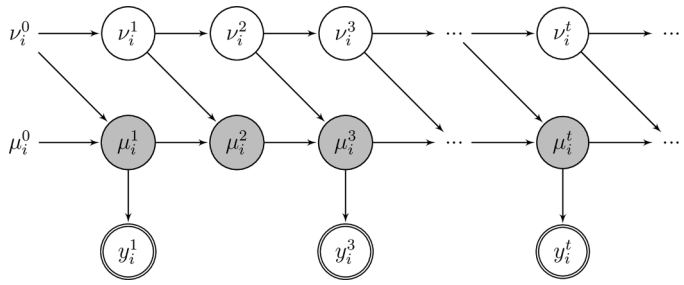


Fig. 1. Diagram of local linear trend model.
The local linear trend model consists of two hidden state variables (μ, the level, and ν, the slope) that evolves over time. The latent sSDMT scores (levels) are highlighted in gray, which are measured with error ε. The observed or measured modes are marked with a double circle. Note that the model naturally handles missing values, as it does not require measurements for each time step. Abbreviations: sSDMT, smartphone Symbol Digit Modalities Test.

rest, or the environment. At each time step t, the latent sSDMT state is assumed to change according to the following set of equations (the dynamics of the LLTM):

$$\mu_i^{t+1} = \mu_i^t + \nu_i^t + \xi_i^t$$

$$\nu_i^{t+1} = \nu_i^t + \zeta_i^t$$

The latent state consists of its level μ_i^t and slope ν_i^t. The level term models how the latent sSDMT state changes over time, while the slope term models its rate of change. The change in state level at each time point is given by the current slope plus random patient-specific zero-mean Gaussian noise ξ_i^t. The slope follows a random walk with patient-specific zero-mean Gaussian noise ζ_i^t. Thus, fitting this model to the data of a specific patient involves estimating three parameters: the irregular (residual) variance ($\mathrm{Var}[\varepsilon_i^t]$), the level variance ($\mathrm{Var}[\xi_i^t]$), and the slope variance ($\mathrm{Var}[\zeta_i^t]$). Concurrently, the best-fitting (maximum likelihood) latent state value is estimated at each time point. This de-noised sSDMT is referred as the estimated sSDMT score.

2.4. Statistical analysis

2.4.1. Measurement error

Reproducibility of measured sSDMT scores was quantified by calculating the standard error of measurement (SEM) from test-retest scores: $\mathrm{SEM} = \mathrm{SD}_{\mathrm{pooled}} \times \sqrt{1 - \mathrm{ICC}}$, where $\mathrm{SD}_{\mathrm{pooled}} = \sqrt{\frac{\mathrm{SD}_{\mathrm{test}}^2 + \mathrm{SD}_{\mathrm{retest}}^2}{2}}$ and ICC is the intraclass correlation coefficient (one-way random, absolute agreement, single-measure). (de Vet et al., 2011; Koo and Li, 2016) With most practice effects within the first two months, for each patient the first sSDMT score from week 9 onwards and a second score within one week were used as test-retest scores as no real change was assumed in this short interval. The SDC was then calculated: $\mathrm{SDC} = 1.96 \times \sqrt{2} \times \mathrm{SEM}$. (de Vet et al., 2011) For the estimated sSDMT, the uncertainty of the LLTM in estimating the latent state, the so-called smoothed variance, (Durbin and Koopman, 2012) at time t is equal to $\mathrm{Var}[\mu_i^t|y_i]$, where the vector y_i consists of all measured sSDMT scores for patient i. The LLTM is fitted with the Kalman filter and smoother from which the smoothed variance is also obtained. (Durbin and Koopman, 2012) This smoothed variance is typically smaller than the measurement error. Furthermore, by conditioning on more observations the smoothed variance generally further decreases due to the law of total variance. (Ross, 2010) This results in an estimated state variable that is less affected by noise. The 95% confidence bands for the inferred latent state are then defined as: $\mu_i^t \pm 1.96 \sqrt{\mathrm{Var}[\mu_i^t|y_i]}$. With individual curve-fitting we shift from a group-level analysis, in which patients are grouped and analyzed together, to an individual-level analysis, in which a model is fit separately for each patient.

2.4.2. Responsiveness

Change in measured and estimated sSDMT scores was assessed for responsiveness to clinically relevant change in 3-month cSDMT (i.e. ≥4 point change (Benedict et al., 2017)) and 12-month BICAMS (i.e. change of ≥4 points in cSDMT, or ≥0.5 SD change in CVLT-II or BVMT-R (Watt et al., 2021)). Measured sSDMT scores within seven days of the clinical visits were averaged. For the estimated sSDMT, the level scores on the exact clinical visit days were used. For M_{12} the estimated sSDMT score within one week was used, since there were no measurements beyond M_{12} for the LLTM to estimate the level. Patients with clinical improvement were grouped and compared to patients with no improvement (i.e. patients who were stable or worsened). Separately, patients with clinical worsening were compared to patients with no worsening. Since there were repeated observations (multiple 3-month periods) within patients, weighted averages (replacing individual observations by its overall subject mean) were calculated to account for differences in the number of observations between patients and correlated observations within patients. The weighted average changes in sSDMT were then used to calculate its 3-month average responsiveness to the clinical cognition outcomes, similar to the calculation of a weighted correlation coefficient with repeated measurements. (Bland and Altman, 1995)

Responsiveness was quantified using receiver operating characteristic (ROC) curves by plotting the true positive rate (sensitivity) against the false positive rate (1 − specificity) of changes in sSDMT for the detection of clinically relevant change in cognitive outcomes. The area under the ROC curve (AUC) was calculated and values ≥0.70 indicated adequate responsiveness. (Prinsen et al., 2018) The change between M_0–M_3 was omitted from this analysis because of the strong practice effects in this period. The sSDMT change that corresponds with the highest Youden's index (senstivity + specificity − 1), most optimally detects clinically relevant change. (van Munster et al., 2020) This cut-off, an approximation of the minimal clinically important difference (MCID), was determined and compared to the SDC to determine whether this smallest important change can be reliably distinguished from measurement error. (de Vet et al., 2011)

2.4.3. Visualization of estimated sSDMT

Individual sSDMT levels were estimated using the state-space model based on all measured sSDMT scores in individual patients. The estimated levels constitute the curve that best fits all the measured sSDMT scores and can be plotted to visualize the trajectory of the estimated sSDMT level over time along with its 95% confidence interval (CI). To investigate the impact of sSDMT assessment frequency on the 95% CI width (i.e. the number of sSDMT points between the upper and lower bound), five patients were selected with (nearly) daily sSDMT scores for at least six months. While downsampling (i.e. lowering) the sSDMT assessment frequency, the change in CI bandwidth was investigated. For downsampling, every n-th data point was taken and varied between 1 (all data points) and 22 (every 22nd data point).

3. Results

A total of 102 patients with MS were included of whom 6 dropped out of the study before M_3, 1 each before M_6 and M_9, and 3 before M_{12}. Two additional patients were included on top of the planned sample size since inclusion was ongoing when the first two patients dropped out. A total of 9319 sSDMT were completed. The adherence rate, computed as the amount of completed tests as percentage of scheduled tests for all patients (including the eventual dropouts), of the sSDMT was 91.5% in the first 4-week period ($n = 102$), 88.4% between week 5 and M_3 ($n = 102$), 82.7% between M_3 and M_6 ($n = 95$), 77.0% between M_6 and M_9 ($n = 94$), and 69.4% between M_9 and M_{12} ($n = 93$). Of the 9319 tests, 1277 (13.7%) were affected by a software bug that caused a slowed application performance during the test and were excluded, and 10 (0.11%)

tests were excluded due to an error leading to a test duration >90 s. The remaining 8032 sSDMT scores were obtained from 100 of the 102 patients and used in the analyses. The median (IQR) amount of sSDMT scores per patient was 66.0 (40.5–90.5). The median (IQR) time between the first and last obtained sSDMT scores was 379.0 (308.0–420.0) days. The baseline demographical and clinical characteristics are summarized in Table 1. The sSDMT averaged around each clinical visit is shown in Fig. 2 along with the average scores for the clinical cognitive outcomes.

3.1. Measurement error

Using measured sSDMT scores, 91 of the 100 patients had a test score and retest score for the reliability analysis. On average, the test was performed 68.6 days after baseline with 6.9 days between the test and retest. With an ICC of 0.90, the SEM and SDC were 2.94 and 8.15 points, respectively. These estimates were derived assuming that all patients share the same measurement error. Concerning the estimated sSDMT, the LLTM was fitted separately for each patient, resulting in different measurement error estimations for every patient. The average SEM across the patients was 2.90 with an estimated spread of 6.55. This large spread strongly suggests that the assumption of different patients having the same SEM, made in the group-level reliability analysis, is unlikely to hold. The average SDC was 8.04.

3.2. Responsiveness

For the responsiveness of the measured sSDMT, smartphone test scores within 7 days of each clinical visit were averaged. This amounted to an average of 7.7 sSDMT scores at M_0 ($n = 92$), 4.6 scores at M_3 ($n = 73$), 2.9 scores at M_6 ($n = 59$), 2.4 scores at M_9 ($n = 42$), and 1.6 scores at M_{12} ($n = 49$). For the estimated sSDMT, as all available sSDMT scores were The results of the responsiveness analysis are shown in Table 2. AUC-values for the measured and estimated sSDMT in classifying relevant change in clinical cognitive outcomes were <0.70. The optimal sSDMT cut-off scores for distinguishing clinically relevant change (i.e. MCID) in cognitive outcomes, were smaller than what can be measured beyond measurement error (i.e. SDC).

3.3. Visualization of estimated sSDMT

All patients exhibited high day-to-day variation in sSDMT performance as single scores changed significantly over consecutive days. The LLTM was individually fitted to the each patients' measured sSDMT scores to obtain the estimated sSDMT level and 95% confidence bands. On average, the estimated standard deviation of the residual component (equivalent to the SEM) was 2.70 points, which is much larger than the average standard deviation in either the level (0.310 points) or slope (0.031 points) component, indicating that real change in processing speed is smooth but measured with high variability Fig. 3. shows two examples of sSDMT performance during the study: one typical patient with an initial increasing trend followed by a relatively stable trend, and one patient with a slight decreasing trend. The 95% confidence bands of the trend estimates fall completely within the spread of raw measurements. This smoothed state error is much smaller than the measurement error of both the sSDMT and cSDMT. Thus, by considering the estimated sSDMT scores as opposed to raw sSDMT or cSDMT scores, the signal-to-noise ratio is improved and statistical power to detect changes over time is gained.

Additionally, the estimated slope state from the LLTM can be used as indication for duration of practice effects. Initially, the estimated slope of the trend of patient A in Fig. 3 is positive; the sSDMT score is improving, likely due to practice effects. At the dashed vertical line in Fig. 3 is the first time that the confidence bands for the slope contain zero. From that point in time the estimated sSDMT can be described as stable, as there is continuous overlap between the confidence bands of the slope and zero. Practice effects were saturated within two months in the vast majority of patients. In patient B the slope variation was not distinct enough to be distinguished from the general variation (in the level), so in the model fit it is attributed to the level variation. Lastly, when inspecting curve fits at two time points, a change larger than half the 95% CI bandwidth multiplied by $\sqrt{2}$ is a statistically significant change. In Fig. 4, this 'change bandwidth' is visualized as function of varying sSDMT assessment frequency. If, for instance, a 4-point MCID in the sSDMT score is considered, (Benedict et al., 2017) at least one measurement every 12 days is needed to detect clinically relevant change with sufficient confidence.

Table 1
Demographic and clinical characteristics at baseline.

	Patients with MS ($n = 100$)
Age, years, mean (SD)	46.5 (10.3)
Sex, n (%) Female Male	74 (74.0%) 26 (26.0%)
Education, n (%) Low Middle High	3 (3.0) 33 (33.0) 64 (64.0)
MS type, n (%) RRMS SPMS PPMS	60 (60.0) 29 (29.0) 11 (11.0)
Disease duration, years, median (IQR)	5.7 (3.1–27.1)
EDSS, median (IQR)	3.5 (2.5–6.0)

Abbreviations: MS, multiple sclerosis; RRMS, relapsing remitting MS; SPMS, secondary progressive MS; PPMS, primary progressive MS; EDSS, Expanded Disability Status Scale.

4. Discussion

In this one-year follow-up study, we investigated a weekly scheduled self-administered smartphone SDMT. Outside of the first two months, where practice effects were most pronounced, the sSDMT had a SDC of 8.15 points. sSDMT scores were found to be insufficiently responsive to change in cSDMT and BICAMS outcomes. Furthermore, as with the cSDMT, sSDMT measurements showcased large variability, which we hypothesize is not solely due to real changes in processing speed. To better understand and clinically monitor how processing speed develops over time, we employed a linear state space model at the individual patient-level for modeling and visualizing the underlying trend of sSDMT measurements. As the error of the curve fitted trend estimate was much smaller than the measurement error in both the raw sSDMT and cSDMT scores, an improved sensitivity to statistically reliable change on the individual patient-level is achieved.

Although cognitive impairment is a well-recognized symptom in MS, identifying clinically meaningful change in cognition remains an inscrutable problem. (Sumowski et al., 2018) The insufficient responsiveness of the sSDMT to clinical cognitive outcomes can be explained by the large irregular variability in both the smartphone and clinical measurements, as evidenced by the large SDC estimates for both metrics. From previous studies the SDC of the cSDMT was estimated to be 11.5, (Benedict et al., 2008) 11.1, (SA Morrow et al., 2010) and 15.2 points, (Sonder et al., 2014) whereas clinically relevant change was established at just 4 points. (Benedict et al., 2017) Additionally, the studied cognitive domains are fairly stable constructs over a period of 1–2 years in which large changes or worsening are often subtle. This was illustrated in two studies where clinical cognitive measures remained stable or even improved throughout a 2-year follow-up period, despite patients subjectively experiencing cognitive decline. (Skorve et al., 2020; Koch et al., 2021) Similarly, our cohort had slight improvements due to practice effects on group level, whereas the number of worsening occurrences was relatively small. Therefore, lack of responsiveness of the sSDMT may be explained by the subtlety or absence of changes in the cognitive outcomes measured with high measurement error.

This suggests that the clinical cognitive measures itself are also insufficiently responsive to change in cognition, as was recently pointed out. (Weinstock et al., al.) With this absence of a clinical ground truth, we made use of the sSDMT's relative abundance of measurements. By curve fitting the high frequency and highly variable smartphone

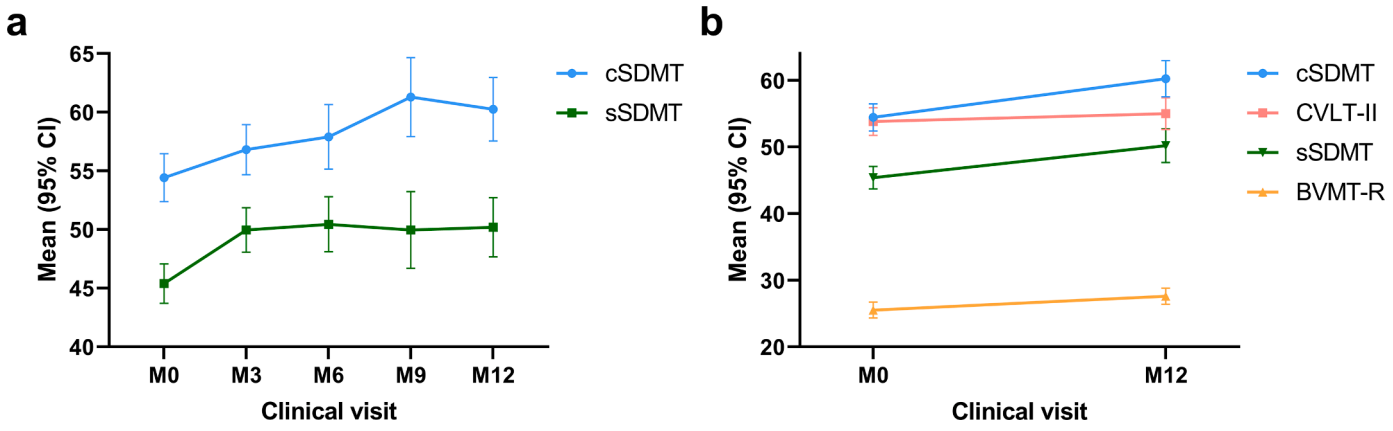


Fig. 2. Line graphs of the smartphone sSDMT and clinical cognitive outcomes.
At 3-month intervals (a), the sSDMT increased significantly between M_0–M_3 (mean difference=5.2 points) and for the cSDMT between M_0–M_3 (2.5 points) and M_6–M_9 (2.9 points). Between M_0–M_{12} (b), mean changes increased significantly for sSDMT (6.5 points), cSDMT (5.6 points), CVLT-II (1.7 points), and BVMT-R (1.8 points).
Abbreviations: cSDMT, clinical Symbol Digit Modalities Test; sSDMT, smartphone Symbol Digit Modalities Test; CVLT-II, California Verbal Learning Test-II; BVMT-R, Brief Visuospatial Memory Test-Revised.

Table 2
Results of responsiveness analysis of sSDMT in classifying relevant change in clinical cognitive outcomes.

	Measured sSDMT			Estimated sSDMT		
	n	AUC (95% CI)	MCID[a]	n	AUC (95% CI)	MCID[a]
3-month cSDMT[b]						
Improvement, yes/no	31 / 74	0.575 (0.444–0.706)	4.67	45 / 94	0.528 (0.415–0.641)	1.22
Worsening, yes/no	21 / 84	0.577 (0.424–0.730)	−0.62	26 / 113	0.512 (0.382–0.642)	−0.13
12-month BICAMS[c]						
Improvement, yes/no	46 / 9	0.667 (0.436–0.898)	2.73	46 / 9	0.594 (0.375–0.813)	0.64
Worsening, yes/no	15 / 40	0.479 (0.291–0.668)	1.48	15 / 40	0.513 (0.330–0.696)	1.72

[a] Most optimal cut-off value based on the maximum Youden's index.
[b] Score change of ≥4 points.
[c] Score change of ≥4 points on SDMT or ≥0.5 SD decrease in CVLT-II or BVMT-R.
Abbreviations: sSDMT, smartphone Symbol Digit Modalities Test; AUC, area under the curve; MCID, minimal clinically important difference; cSDMT, clinical SDMT; BICAMS, Brief International Cognitive Assessment for Multiple Sclerosis; CVLT-II, California Verbal Learning Test-II; BVMT-R, Brief Visuospatial Memory Test-Revised.

measurements, a de-noised trend estimate was derived that more accurately reflects true changes in processing speed and allows visualization over time. We proposed a state space model, which is used to describe complex processes like processing speed by assuming an underlying state space that follows clearly specified dynamics. We settled on a LLTM which includes a varying slope for the underlying trend to account for specific variation in the sSDMT scores such as practice effects. As output, we provided the trend line estimate of the latent sSDMT state with corresponding 95% confidence intervals expressing the uncertainty in trend estimation. A similar approach was proposed for Parkinson's disease, in which a linear state space model was used to model longitudinal data of a disease rating scale to estimate its reliability. (Evers et al., 2019) We showed how the fitted curves can be used to analyze whether a patient's de-noised sSDMT score is truly undergoing changes over time. We also used the fitted parameters from the LLTM to quantify how much variation is due to real change, and the duration of practice effects. Similarly, Pham et al. employed non-linear regression to identify the time point when practice effects stabilize for their app-based SDMT. (Pham et al., 2021)

Curve fitting frequent test scores provides a more robust trend estimate and its 95% CI with a much smaller measurement error compared to singular or clinical tests. This enables the monitoring of cognition with an increased statistical reliability on the individual level. For practical guidance, we would consider this 95% CI bandwidth to indicate significant changes in cognition to complement current clinical monitoring of patients. With the advantages gained from higher-frequency measurements, we analyzed the minimum sSDMT frequency necessary to obtain a sufficiently robust latent state estimate. When downsampling patients with daily sSDMTs, the 95% confidence bands of the estimated sSDMT level changed according to the "square root law" of statistical precision. (Freedman et al., 2007) Based on the derived minimum frequency of once per 12 days, we propose a sSDMT monitoring frequency of (at least) once per week, as this fits more naturally into a weekly routine and promotes adherence. Furthermore, even smaller sSDMT changes may be of interest when combined with other (digital) biomarkers, e.g. for ambulatory or upper limb function, to monitor a multidimensional disease such as MS and improve sensitivity to potentially detect disease activity and progression.

A limitation of our study is that it is based on a general, unselected cohort of MS patients, to investigate the use of smartphone biomarkers in MS for multiple disease outcomes and not specifically for cognition. Another limitation is that 13.8% tests were excluded due to technical problems of the sSDMT app. Other limitations include the assumption that the sSDMT evolution is governed by linear equations. However, the LLTM is flexible enough to capture complex non-stationary sSDMT variations for individual patients. (Durbin and Koopman, 2012) Even if the sSDMT evolution is nonlinear, a piecewise linear approximation is reasonable given a sufficient assessment frequency. Alternatively, a range of nonparametric methods (e.g. LOESS regression) were also compared yielding similar results. While nonparametric approaches require fewer assumptions, the state space model handles time series data more naturally by assuming an underlying "real" sSDMT score

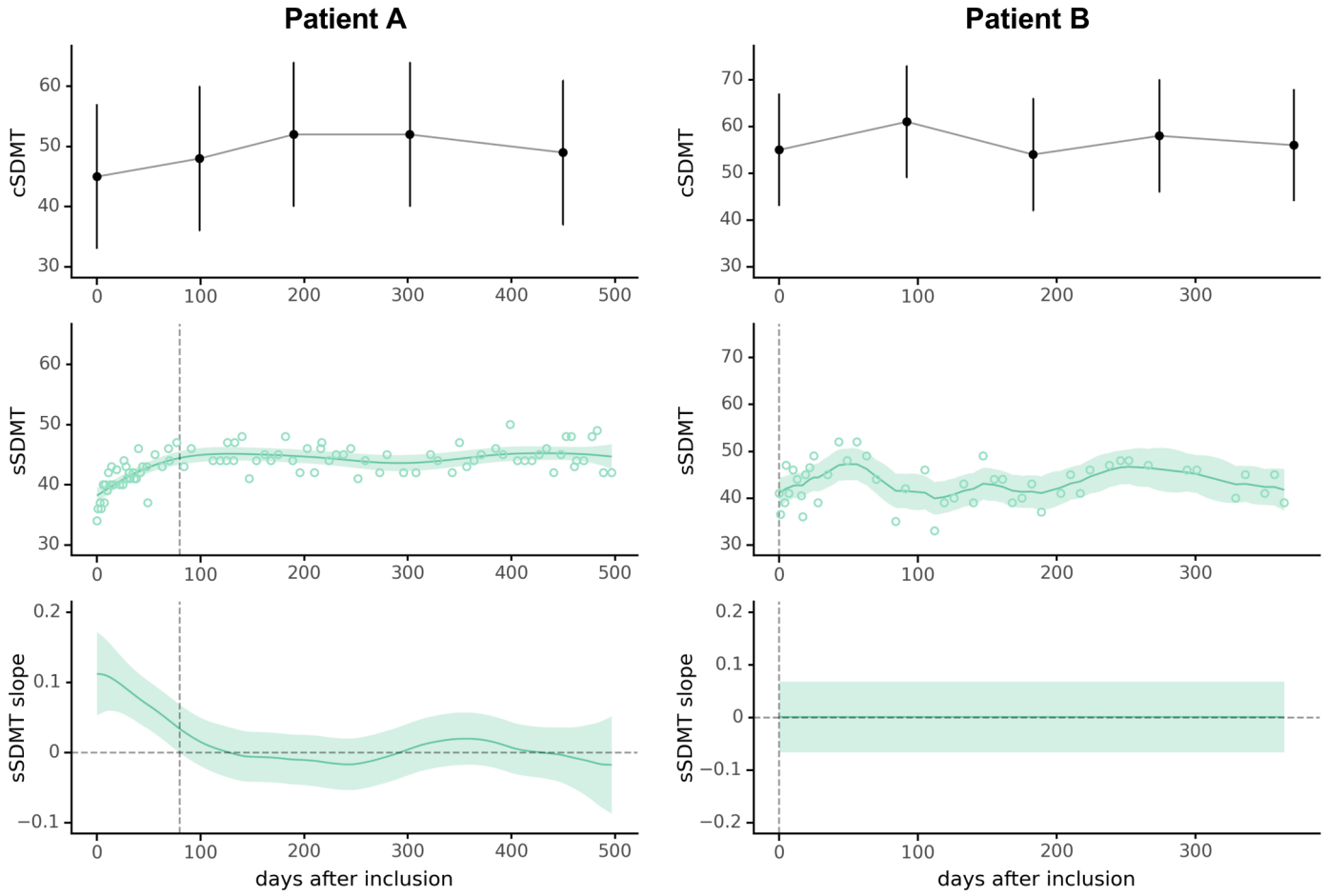


Fig. 3. Curve-fitting examples of the local linear trend model.
The cSDMT scores (dots) and its 12-point smallest detectable change (vertical bars) are shown in the upper panels. The local linear trend model fit (solid line) and 95% CI (band) are superimposed on the measured sSDMT scores (circles) in the middle panels. The lower panels shows the mean slope estimate (solid line) and 95% CI (band) of the trend. The vertical dashed line in patient A represents the cut-off time point when the confidence band of the slope estimate intersects zero on the y-axis, indicating the (estimated) time when the practice effect stabilizes.
Abbreviations: cSDMT, clinical Symbol Digit Modalities Test; sSDMT, smartphone Symbol Digit Modalities Test.

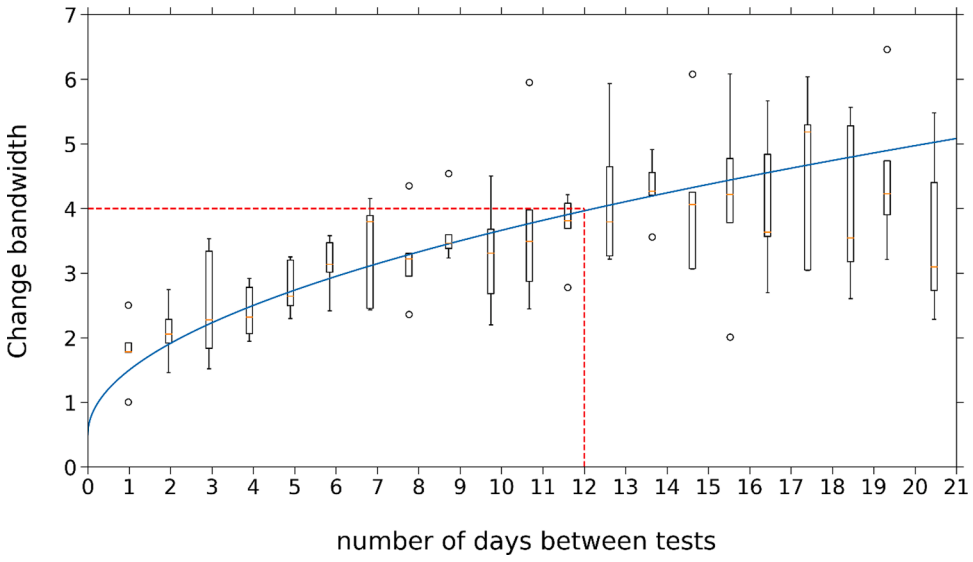


Fig. 4. Box plots visualizing the change bandwidth in five selected patients with the highest sSDMT assessment frequency.
The change in bandwidth size of the sSDMT model (y-axis) was estimated as a function of assessment frequency (x-axis). An expected "square root of d" behavior was observed (solid curve), where d is the average number of days in between tests; as the number of measurement increases by downsampling, the change bandwidth decreases roughly by a factor of $\sqrt{d}$.
Abbreviations: sSDMT, smartphone Symbol Digit Modalities Test.

exhibiting day-to-day variability in observed measurements. Additionally, state space models enable variation quantification through fitted parameters. Another limitation is when performing the LLTM fit in patients with few measurements, the maximum likelihood estimation can get stuck in a poor local optimum. While we show that the assumptions underlying a group-level analysis are unlikely to hold, a possible solution could be to perform a multi-level Bayesian analysis with "soft sharing" of parameters between patients (as opposed to "hard sharing" in group-level analysis) by using a common distribution. However, this approach would increase the complexity of the estimation. We leave its analysis for future work.

5. Conclusions

A self-administered smartphone SDMT was shown to have improved reproducibility estimates compared to the clinical SDMT due to higher measurement frequency. Despite this, the sSDMT was unresponsive to clinical cognitive outcomes as a result of high variability in scores and, even more so, high variability in clinical cognitive outcomes. To circumvent the high variability, a curve fitting approach making use of the high frequency smartphone tests was derived to estimate a de-noised sSDMT score. This curve fitting approach allows visualization of the sSDMT score trajectory and its confidence bands for individual patients. For clinical monitoring, with weekly measurements the estimated confidence bands of the trend can be used to indicate significant changes in cognition. Therefore, this visualization approach enables fine-grained individual-based monitoring that improves the detection of statistically reliable change in MS.

Funding

The authors disclosed receipt of the following financial support for the research: the collaboration project was co-funded by the PPP Allowance made available by Health~Holland, Top Sector Life Sciences & Health [grant number LSHM16060-SGF] and Stichting MS Research [grant number 16–946 MS] to stimulate public-private partnerships, and by a contribution from Biogen (unrestricted funding). The collaboration with the Institute for Computing and Information Sciences (Radboud University) was made possible by funding from the Dutch Research Council (NWO) and Nationaal MS fonds.

Data availability

Anonymized data not published within the article is available upon request from a qualified investigator. Such requests must be submitted in writing and will be reviewed regarding criteria for researcher qualifications and legitimacy of the research purpose.

CRediT authorship contribution statement

Ka-Hoo Lam: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Ioan Gabriel Bucur:** Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Pim Van Oirschot:** Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Frank De Graaf:** Validation, Formal analysis, Writing – review & editing. **Hans Weda:** Validation, Formal analysis, Writing – review & editing. **Eva Strijbis:** Writing – review & editing. **Bernard Uitdehaag:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Funding acquisition. **Tom Heskes:** Writing – review & editing, Supervision. **Joep Killestein:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Vincent De Groot:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

K.H. Lam, I.G. Bucur, E.M.M. Strijbis, T. Heskes, and V. de Groot have no conflicts of interest. P. van Oirschot, F. de Graaf, and H. Weda are employees of Orikami Digital Health Products (industry partner). B. M.J. Uitdehaag received consultancy fees from Biogen Idec, Genzyme, Merck Serono, Novartis, Roche, and Teva. J. Killestein has accepted speaker and consultancy fees from Merck, Biogen, Teva, Genzyme, Roche, and Novartis.

Acknowledgements

The authors would like to express their gratitude to all the patients who participated in the study.

References

Benedict, R.H., DeLuca, J., Phillips, G., LaRocca, N., Hudson, L.D., Rudick, R., 2017. Validity of the Symbol Digit Modalities Test as a cognition performance outcome measure for multiple sclerosis. Mult. Scler. 23 (5), 721–733.
Benedict, R.H., Duquin, J.A., Jurgensen, S., Rudick, R.A., Feitcher, J., Munschauer, F.E., et al., 2008. Repeated assessment of neuropsychological deficits in multiple sclerosis using the Symbol Digit Modalities Test and the MS Neuropsychological Screening Questionnaire. Mult. Scler. 14 (7), 940–946.
Benedict, R.H.B., Amato, M.P., DeLuca, J., Geurts, J.J.G., 2020. Cognitive impairment in multiple sclerosis: clinical management, MRI, and therapeutic avenues. The Lancet Neurology 19 (10), 860–871.
Benedict, R.H., Pol, J., Yasin, F., Hojnacki, D., Kolb, C., Eckert, S., et al., 2021. Recovery of cognitive function after relapse in multiple sclerosis. Mult. Scler. 27 (1), 71–78.
Benedict, R.H.B., Smerbeck, A., Parikh, R., Rodgers, J., Cadavid, D., Erlanger, D., 2012. Reliability and equivalence of alternate forms for the Symbol Digit Modalities Test: implications for multiple sclerosis clinical trials. Multiple Sclerosis Journal 18 (9), 1320–1325.
Bland, J.M., Altman, D.G., 1995. Calculating correlation coefficients with repeated observations: part 2–Correlation between subjects. BMJ (Clinical research ed) 310 (6980), 633.
de Vet, H.C.W., Terwee, C.B., Mokkink, L.B., Knol, D.L., 2011. Measurement in Medicine: A Practical Guide. Cambridge University Press, Cambridge.
Drake, A.S., Weinstock-Guttman, B., Morrow, S.A., Hojnacki, D., Munschauer, F.E., Benedict, R.H., 2010. Psychometrics and normative data for the Multiple Sclerosis Functional Composite: replacing the PASAT with the Symbol Digit Modalities Test. Mult. Scler. 16 (2), 228–237.
Durbin, J., Koopman, S.J., 2012. Time Series Analysis By State Space Methods, 2nd Edition. OUP Oxford, p. 369. 2012/05/03/p.
Evers, L.J.W., Krijthe, J.H., Meinders, M.J., Bloem, B.R., Heskes, T.M., 2019. Measuring Parkinson's disease over time: the real-world within-subject reliability of the MDS-UPDRS. Mov. Disord. 34 (10), 1480–1487.
Freedman, D., Pisani, R., Purves, R., 2007. Statistics, 4th edition. WW Norton, New York.
Giedraitiene, N., Kaubrys, G., Kizlaitiene, R., 2018. Cognition During and After Multiple Sclerosis Relapse as Assessed With the Brief International Cognitive Assessment for Multiple Sclerosis. Sci. Rep. 8 (1), 8169.
Kalb, R., Beier, M., Benedict, R.H., Charvet, L., Costello, K., Feinstein, A., et al., 2018. Recommendations for cognitive screening and management in multiple sclerosis care. Mult. Scler. 24 (13), 1665–1680.
Koch, M.W., Mostert, J., Repovic, P., Bowen, J.D., Uitdehaag, B., Cutter, G., 2021. Is the Symbol Digit Modalities Test a useful outcome in secondary progressive multiple sclerosis? Eur. J. Neurol.
Koo, T.K., Li, M.Y., 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 15 (2), 155–163.
Kurtzke, J.F., 1983. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 33 (11), 1444–1452.
Lam, K.H., van Oirschot, P., den Teuling, B., Hulst, H.E., de Jong, B.A., Uitdehaag, B., et al., 2021. Reliability, construct and concurrent validity of a smartphone-based cognition test in multiple sclerosis. Mult. Scler., 13524585211018103
Langdon, D.W., Amato, M.P., Boringa, J., Brochet, B., Foley, F., Fredrikson, S., et al., 2012. Recommendations for a Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS). Mult. Scler. 18 (6), 891–898.
Maillart, E., Labauge, P., Cohen, M., Maarouf, A., Vukusic, S., Donze, C., et al., 2020. MSCopilot, a new multiple sclerosis self-assessment digital solution: results of a comparative study versus standard tests. Eur. J. Neurol. 27 (3), 429–436.
Morrow, S.A., Drake, A., Zivadinov, R., Munschauer, F., Weinstock-Guttman, B., Benedict, R.H., 2010a. Predicting loss of employment over three years in multiple sclerosis: clinically meaningful cognitive decline. Clin. Neuropsychol. 24 (7), 1131–1145.
Morrow, S.A., O'Connor, P.W., Polman, C.H., Goodman, A.D., Kappos, L., Lublin, F.D., et al., 2010b. Evaluation of the symbol digit modalities test (SDMT) and MS neuropsychological screening questionnaire (MSNQ) in natalizumab-treated MS patients over 48 weeks. Mult. Scler. 16 (11), 1385–1392.

Netherlands Trial Register. Trial NL7070 Assessing fatigue, disease activity and progression through smartphone surveillance in multiple sclerosis 2018 [cited 2022 January 24]. Available from: https://www.trialregister.nl/trial/7070.

Orikami Digital Health Products. MS sherpa® [Internet]. [cited 2021 April 8]. Available from: https://www.mssherpa.com.

Pham, L., Harris, T., Varosanec, M., Morgan, V., Kosa, P., Bielekova, B., 2021. Smartphone-based symbol-digit modalities test reliably captures brain damage in multiple sclerosis. NPJ digital medicine 4 (1), 36.

Prinsen, C.A.C., Mokkink, L.B., Bouter, L.M., Alonso, J., Patrick, D.L., de Vet, H.C.W., et al., 2018. COSMIN guideline for systematic reviews of patient-reported outcome measures. Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation 27 (5), 1147–1157.

Pv, Oirschot, Heerings, M., Wendrich, K., Bd, Teuling, Martens, M.B., Jongen, P.J., 2020. Symbol Digit Modalities Test Variant in a Smartphone App for Persons With Multiple Sclerosis: validation Study. JMIR Mhealth Uhealth 8 (10), e18160.

Rijnen, S.J.M., Meskal, I., Emons, W.H.M., Campman, C.A.M., van der Linden, S.D., Gehring, K., et al., 2020. Evaluation of Normative Data of a Widely Used Computerized Neuropsychological Battery: applicability and Effects of Sociodemographic Variables in a Dutch Sample. Assessment 27 (2), 373–383.

Ross, S.M., 2010. A First Course in Probability. Pearson Prentice Hall.

Skorve, E., Lundervold, A.J., Torkildsen, Ø., Myhr, K.-.M., 2020. A two-year longitudinal follow-up of cognitive performance assessed by BICAMS in newly diagnosed patients with MS. Mult Scler Relat Disord 46, 102577.

Sonder, J.M., Burggraaff, J., Knol, D.L., Polman, C.H., Uitdehaag, B.M., 2014. Comparing long-term results of PASAT and SDMT scores in relation to neuropsychological testing in multiple sclerosis. Multiple Sclerosis Journal 20 (4), 481–488.

Sumowski, J.F., Benedict, R., Enzinger, C., Filippi, M., Geurts, J.J., Hamalainen, P., et al., 2018. Cognition in multiple sclerosis: state of the field and priorities for the future. Neurology 90 (6), 278–288.

Thompson, A.J., Banwell, B.L., Barkhof, F., Carroll, W.M., Coetzee, T., Comi, G., et al., 2018. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. The Lancet Neurology 17 (2), 162–173.

van Munster, C.E., Kaya, L., Obura, M., Kalkers, N.F., Uitdehaag, B.M., 2020. Minimal clinically important difference of improvement on the Arm Function in Multiple Sclerosis Questionnaire (AMSQ). Mult. Scler. 26 (4), 505–508.

Watt, J.A., Veroniki, A.A., Tricco, A.C., Straus, S.E., 2021. Using a distribution-based approach and systematic review methods to derive minimum clinically important differences. BMC Med. Res. Methodol. 21 (1), 41.

Weinstock Z., Morrow S., Conway D., Fuchs T., Wojcik C., Unverdi M., et al. Interpreting change on the Symbol Digit Modalities Test in people with relapsing multiple sclerosis using the reliable change methodology. Multi. Scler. Journ. 0(0): 13524585211049397.