MSJ Experimental Translational Clinical

*Original Research Paper*

# Multicentre assessment of motor and sensory evoked potentials in multiple sclerosis: reliability and implications for clinical trials

*Multiple Sclerosis Journal—Experimental, Translational and Clinical*

April–June 2019, 1–11

DOI: 10.1177/2055217319844796



**Martin Hardmeier, François Jacques, Philipp Albrecht, Habib Bousleiman, Christian Schindler, Letizia Leocani, Peter Fuhr**

## *Abstract*

**Background:** Motor and sensory evoked potentials (EP) are potential candidate biomarkers for clinical trials in multiple sclerosis.

**Objective:** To determine test -retest reliability of motor EP (MEP) and sensory EP (SEP) and associated EP-scores in patients with multiple sclerosis.

**Methods:** In three centres, 16 relapsing and five progressive multiple sclerosis patients had MEPs and SEPs 1–29 days apart. Five neurophysiologists independently marked latencies by central reading. By variance component analysis, we estimated the critical difference (absolute reliability) for cross-sectional group comparison, comparison of longitudinal group changes, within-subject minimal detectable change and defined within-subject improvement.

**Results:** Cortical SEP responses and cortico-muscular MEP latencies were more reliable than central conduction times. For comparison of 20 subjects per arm, cross-sectional group difference ranged from 0.7 to 3.9 ms and 1.1 to 1.7, group difference in longitudinal changes from 0.4 to 1.8 ms and 0.36 to 0.62, within-subject minimal detectable change from 1.2 to 5.8 ms and 1.2 to 2.0, within-subject improvement from 0.8 to 3.8ms and 0.8 to 1.3, for single EP modalities and EP scores, respectively.

**Conclusions:** Multicentre EP assessment with central EP reading is feasible and reliable. The critical difference is reasonably low to detect significant group changes and to define responders. The results support the concept of using EP and EP-scores as candidate response biomarkers for quantification of disease progression and for studying remyelination in multiple sclerosis.

***Keywords:*** Motor evoked potentials, sensory evoked potentials, biomarker, response biomarker, progressive MS, remyelination, test–retest reliability, mean detectable change

Date received: 21 November 2018; accepted: 23 March 2019

## Introduction

For clinical trials in multiple sclerosis (MS) targeting disease progression and remyelination there is still a need for surrogate outcomes. Such response biomarkers have to be valid, reliable and sensitive to change.[1] Evoked potentials (EPs) have a high construct and criterion validity[2] reflected by their close relationship to the pathophysiology of symptoms,[3,4] and by the fact that EP scores derived from different EP modalities closely correlate to clinical disability and predict clinical course over up to 20 years.[5–11] In addition, EPs are unbiased for directional change, while remaining specific for the neuronal function, and thus can measure deterioration as well as improvement. In the current study, we aimed to determine the reliability of quantitative EP scores as candidate response biomarkers for clinical trials in MS.

Correspondence to:
**Peter Fuhr,**
Department of Neurology, University Hospital Basel, Petersgraben 4, 4031 Basel, Switzerland.
**peter.fuhr@usb.ch**

**Martin Hardmeier**, Department of Neurology, Hospital of the University of Basel, Switzerland

**François Jacques**, Neurology, Clinique Neuro-Outaouais, Canada

**Philipp Albrecht**, Department of Neurology, Heinrich Heine University Düsseldorf, Germany

**Habib Bousleiman**, Department of Neurology, Hospital of the University of Basel, Switzerland

**Christian Schindler**, Swiss Tropical and Public Health Institute, University of Basel, Switzerland

**Letizia Leocani**, Departments of Neurology and Neurorehabilitation, Ospedale San Raffaele, Milano, Italy

**Peter Fuhr**, Department of Neurology, Hospital of the University of Basel, Switzerland

To define the capability of a measure for detecting a real within-subject change it is important to differentiate relative from absolute reliability.[12,13] Relative reliability denotes the stability of a value's rank over time relative to the entire sample and is valuable for diagnostic and prognostic purposes. Absolute reliability depends on the measurement error within the same subject and is important for defining the minimal detectable change (MDC) in a test–retest situation.[14,15]

In most studies on measures derived from motor evoked potentials (MEPs) relative reliability is reported.[13] The absolute reliability of MEPs is higher for latencies than amplitudes, and MDCs in single subjects range from 0.7 to 3.5 ms in hand muscles and from 1.8 to 1.9 ms in a leg muscle.[16–18] In a large multicentre study, the mean within-subject standard deviation (SD) was 0.85 ms in a hand muscle and the relative inter-session variability was below 10% for a leg muscle.[19,20] In patients with relapsing and chronic MS, the within-subject intra-session variability was increased in 67% of patients as compared to healthy subjects.[21]

Reliability studies in sensory evoked potentials (SEPs) are scarce and old reports include less than 10 subjects in serial recordings.[22,23] A large recent trial showed high reliability for the N20 latency in median SEPs with a mean within-subject SD of 0.48 ms.[19] No recent study has reported on the reliability of tibial SEPs.

The reliability of the combination of different EP modalities, i.e. EP scores, has not been studied. In the case of MS, EP scores better reflect the extent of impaired signal propagation than single modalities due to the disseminated pathology of MS, and long tracts have a higher probability to be altered. Numerical scores calculated from z-transformed EP latencies have been reported to have higher sensitivity to change than ordinal or semiquantitative scores;[24] however, they may be prone to higher measurement variability.

In the current three centre study, we addressed the measurement variability of SEPs, MEPs and EP scores in a group of patients with relapsing–remitting, secondary and primary progressive MS. We employed a novel standardised recording protocol, a custom-made server-based software for standardised curve reading (EPMark) and a statistical variance component analysis. We determined the critical difference for a cross-sectional group comparison, for comparison of longitudinal changes between two groups, and the MDC and, furthermore, suggest a definition of a within-subject improvement.

## Participants and methods

### *Participants*

The study was approved by the local institutional review boards at the three participating study sites (Basel, Dusseldorf, Gatineau) and was conducted according to International Conference on Harmonisation Good Clinical Practice (ICH-GCP) guidelines. All participants gave written informed consent. Inclusion criteria comprised a diagnosis of MS (all phenotypes), an Expanded Disability Status Scale (EDSS) between 0 and 6.5 and absence of a relapse for at least 3 months. Exclusion criteria included change in medications which possibly interfere with signal propagation (sodium or potassium channel antagonists and spasmolytics), comorbidities, which may affect testing (polyneuropathy, cervical stenosis and morbid obesity among others) and contraindication to MEP recording (epilepsy, movable metal implants).

### *Recording*

Patients were recorded twice at the same centre with an interval of one to 30 days. Recording of single EP modalities followed closely the recommendations of the International Federation of Clinical Neurophysiology.[25–27] The protocol was optimised for fast and robust acquisition to be feasible in a multicentre setting and the montage for SEPs, as well as the placement of electrodes, coil size and level of pre-innervation of the target muscle for MEPs were standardised (for details see Supplementary file 1). The recording of one limb in one modality is referred to as a 'test'; hence, a subject has a set of eight tests per time point.

### *Rating of curves*

Curves were coded and uploaded to a custom server-based software application (EPMark, Supplementary file 2), which displays curves in a standardised fashion for central reading. All curves were evaluated by experienced neurophysiologists (PA, PF, MH, FJ, LL) blinded to clinical details, and markers were set manually for cortico-muscular (CxM) and spino-muscular (SpM) latencies in MEPs and for peaks of the main cortical (N20, P40), cervical (N13) and lumbar (N22) responses as well as Erb in SEPs. Follow-up rating was done blinded to baseline results after a delay of at least one week.

Curves were excluded if less than three (out of five) raters considered the N20 or the P40 peak or the onset latency as valid. In addition, a MEP test was only included if at least three (out of eight to 10) MEP curves were valid.

### *Definition of EP parameters and quantitative EP scores*

Central conduction is the most important EP measure for diagnostic purposes in MS but is possibly prone to more variability as it is the difference of two measured values (overall and peripheral conduction times). At first-level analysis we therefore evaluated the reliability of both, central and overall latencies as well as two different approaches to define MEP onset latency. At second-level analysis, the most reliable single EP tests were aggregated to quantitative EP scores to yield one-dimensional measures for statistical analysis.

In median SEPs (SEP-M), the cortical N20 and central conduction times (CCTs; CCT-M1 = latency (N20–N13), CCT-M2 = latency (N20–Erb)) were used for statistical analysis, in tibial SEPs (SEP-T) the cortical P40 and the CCT-T (CCT-T = latency (P40–N22)). In MEPs the shortest CxM (shortest latency of at least three curves), mean CxM (mean of at least four curves), and central motor conduction time (CMCT; shortest CMCT = shortest CxM-shortest SpM; mean CMCT = mean CxM-shortest SpM) were analysed for upper limbs (ULs) and lower limbs (LLs).

The method to calculate quantitative EP scores has been described previously.[5,8] Scores equal the sum of z-transformed latencies of each included test divided by the number of tests. They are in the z-space and dimensionless. In the current study, we calculated the EP score without visual evoked potentials (VEPs) and included N20, P40 and CxM instead of CCT and CMCT, as the former show higher reliability at first-level analysis. Two versions of the modified quantitative EP score (mqEPS) were evaluated, the mqEPS-short comprises: z-N20, z-P40, shortest z-CxM from MEP-UL and MEP-LL, the mqEPS-mean: z-N20, z-P40, mean z-CxM from MEP-UL and MEP-LL. Accordingly, we calculated scores based on MEPs (qMEP-short, qMEP-mean) and on SEPs (qSEP). For transformation in z-space, published normative values were used (Supplementary file 3a).

### *Statistical analysis*

To describe the concordance for test and retest, we calculated Spearman's correlation coefficients for each rater and side in MEPs and SEPs, and for each rater in quantitative EP scores. To explore data distribution, we used Bland–Altman plots and inspected them visually.[28]

We used linear mixed effects models with the results of test and retest assessment as the combined outcome and included random factors to estimate the different variance components. For single EP modalities, the model comprised nine random factors (rater, subject; subject*time, subject*rater, subject*side, rater*time (two-way interactions); subject*rater*time; subject*side*rater; subject*side*time (three-way interactions) and an error term (i.e. subject*side*rater*time). As EP scores already include the factor 'side', the corresponding models comprise six factors. The factor 'site' was not considered as only three sites participated rendering the estimate of site variability unreliable.

We calculated the standard error (SE) for a cross-sectional comparison between two groups ($SE_{cross}$), for a comparison of mean longitudinal change between two groups ($SE_{long}$) assuming equal sized groups with equal data distribution. In addition, SE was calculated for a longitudinal within-subject comparison ($SE_{1S}$). The case of two raters was assumed who independently mark all curves of all subjects corresponding to a central reading in a study setting (without consensus reading). Variance components and the weights for defining the three SEs, which were calculated as the square root of the respective sum of variance components are given in Supplementary file 3b.

The critical difference $D_0$ is given as the product of SE by the respective quantile of the standard normal distribution $z_\alpha$ defined by the selected alpha level, yielding $D_p = SE*z_p$. $D_0$ is the value distinguishing values of D compatible with the null hypothesis of no real change from values of D where there is evidence for a true change.

We assumed a two-sided testing with an alpha level of 5% for group comparisons yielding $D_{cross} = SE_{cross}*1.96$ and $D_{long} = SE_{long}*1.96$. For change within a single subject we applied the definition of the MDC[14] giving $D_{MDC} = SE_{1S}*1.96$. To increase sensitivity, we additionally calculated a critical difference for improvement (or progression), for which we considered a one-sided test at an alpha level of 10% as justified, yielding $D_{imp} = SE_{1S}*1.24$.

## Results

Twenty-two subjects were recruited (Basel: nine, Dusseldorf: five, Gatineau: eight), one had no

follow-up exam and was excluded from analysis. The remaining subjects had a median age of 51.0 years (range 20.4–65.3) and 43% were men; 76.2% ($n$=16) had a relapsing–remitting, 14.3% ($n$=3) a secondary progressive and 9.5% ($n$=2) a primary progressive disease course. Different disease courses were included to cover the full range from normal to severely pathological EPs which are more likely in progressive patients. The median disease duration was 9.2 years (range 0.2–40.3), median EDSS 3.0 (range 0–6.5) and last relapse at least a year ago except in three patients (relapse between 80 and 93 days prior to baseline). Disease-modifying and symptomatic treatments remained unchanged during the study.

**Table 1.** Number of tests per modality used for analysis, left and right sides combined.

| | Baseline ($n=22$) | Follow-up ($n=21$) |
|---|---|---|
| MEP-UL | 44 | 40 |
| MEP-LL | 42 | 40 |
| SEP-M | 44 | 42 |
| SEP-T[a] | 39 | 37 |

UL: upper limb; LL: lower limb; M: median; T: tibial.
[a]One subject did not take part in SEP-T.

The median time between recordings was 8 days (range 1–29 days). Table 1 shows the number of tests per modality used for analysis. One subject refused recording of tibial SEPs.

Fifty-nine per cent of subjects had at least one pathological test (one to two path. tests: 22.7%; three to four path. tests: 27.3%; more than four path. tests: 9.1%). SEP-T had the highest yield of pathology (P40: 51.3%, CCT-T: 42.9%), then MEP-LL (shortest CxM: 28.6%, shortest CMCT: 33.3%) and MEP-UL (shortest CxM: 22.7%, shortest CMCT: 29.6%); SEP-M had only a few pathologies in CCT (N20: 0%, CCT-M: 6.8%).

**Table 2.** Spearman's correlation coefficient for the association between test and re-test in single EP modalities and EP scores.

| (a) | MEP-UL | | | | MEP-LL | | | |
|---|---|---|---|---|---|---|---|---|
| | CxM | | CMCT | | CxM | | CMCT | |
| | Shortest | Mean | Shortest | Mean | Shortest | Mean | Shortest | Mean |
| Median | 0.85 | 0.92 | 0.75 | 0.79 | 0.87 | 0.93 | 0.73 | 0.71 |
| Min | 0.80 | 0.91 | 0.65 | 0.75 | 0.76 | 0.86 | 0.65 | 0.60 |
| Max | 0.90 | 0.94 | 0.79 | 0.85 | 0.89 | 0.94 | 0.78 | 0.79 |

| (b) | SEP-M | | | SEP-T | |
|---|---|---|---|---|---|
| | N20 | CCT-M1 | CCT-M2 | P40 | CCT-T |
| Median | 0.92 | 0.54 | 0.73 | 0.79 | 0.74 |
| Min | 0.83 | 0.44 | 0.70 | 0.77 | 0.63 |
| Max | 0.93 | 0.67 | 0.80 | 0.81 | 0.77 |

| (c) | Multimodal qEPS | | Unimodal qEPS | | |
|---|---|---|---|---|---|
| | mqEPS-short | mqEPS-mean | qMEP-short | qMEP-mean | qSEP |
| Median | 0.91 | 0.92 | 0.93 | 0.94 | 0.80 |
| Min | 0.86 | 0.84 | 0.83 | 0.89 | 0.78 |
| Max | 0.96 | 0.95 | 0.96 | 0.97 | 0.83 |

Spearman's rho was determined in each rater and side separately, and median and range is given for single EP modalities and for quantitative EP scores ((a) MEP upper (UL) and lower limb (LL), cortico-muscular latency (CxM) and central motor conduction time (CMCT) determined from shortest or the mean of MEP curves of one test; (b) median (SEP-M) and tibial SEP (SEP-T) with central conduction time (CCT) determined from N20-CV7 (M1) or N20-EP (M2); (c) modified quantitative EP score (mqEPS) and quantitative MEP score (qMEP) from shortest and mean MEP curves, quantitative SEP score (qSEP)).

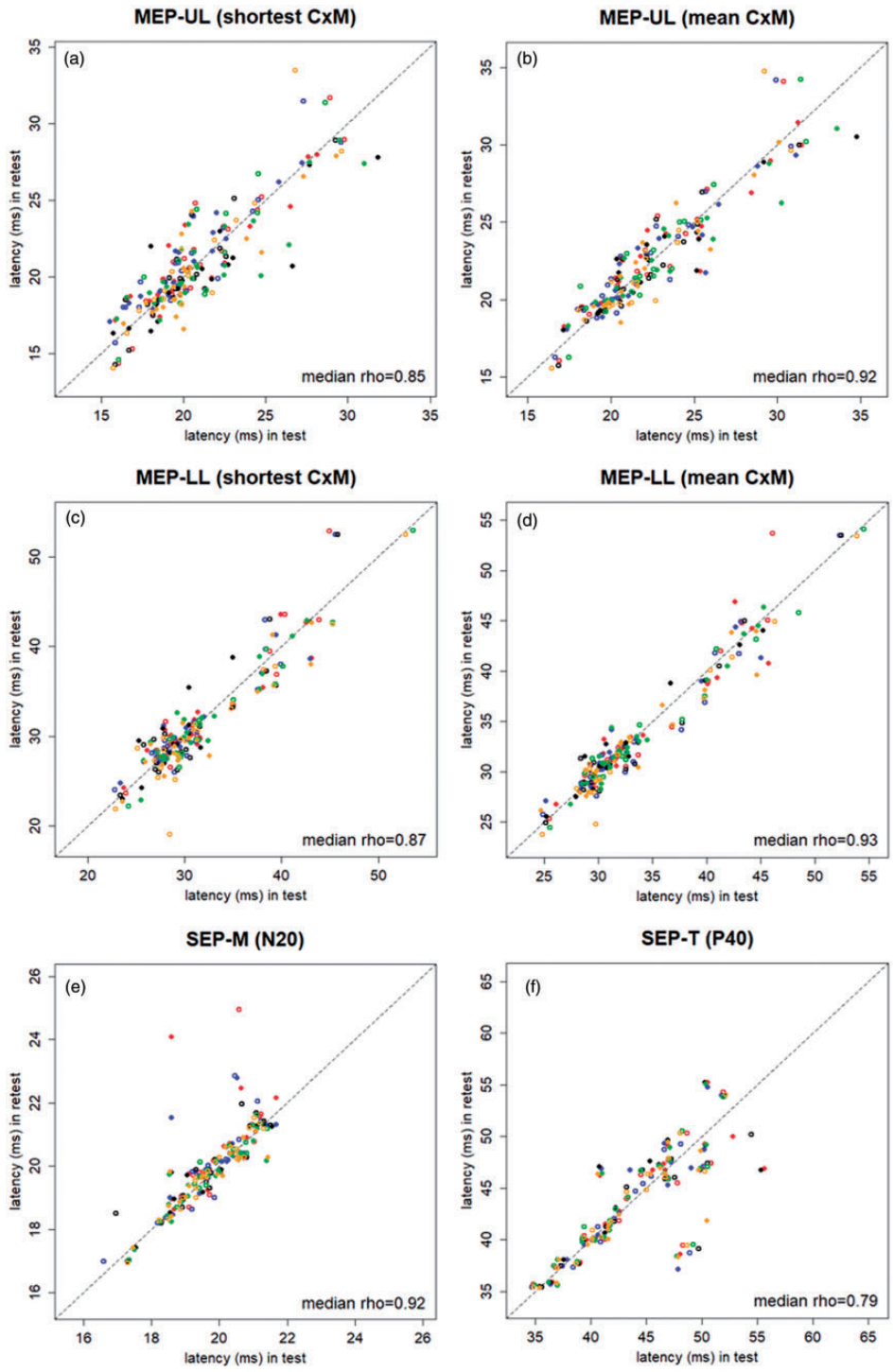


**Figure 1.** Each scatterplot gives an overall impression of the association between test and retest rating in each evoked potential (EP) modality pooled across sides (circles: left; points: right) and raters (colours) ((a) shortest and (b) mean cortico-muscular latency (CxM) for upper limb motor EPs (MEP-UL); (c) shortest CxM and (d) mean CxM for lower limb MEPs (MEP-LL); (e) N20 for median SEPs (SEP-M); (f) P40 for tibial SEPs (SEP-T). Diagonal lines represent perfect concordance between test and retest. Correlational analysis was run on each rater and side separately (2 _ 5 times per measure), median Spearman's rho correlation coefficient is given (please refer to Table 2 and to Supplementary Figure 1 for further details).

### *Correlational analysis*

Table 2 gives the median (range) rho of all readers and sides; Figure 1(a–f) and Supplementary Figure 1 show the related scatterplots. In MEPs, CxM showed higher concordance as compared to CMCT, and mean CxM higher concordance than the shortest CxM. In SEPs, N20 and P40 showed higher concordance than CCT-M1, CCT-M2 and CCT-T. The concordance of EP scores was generally high except for qSEP, and slightly higher in scores based on mean CxM compared to shortest CxM. Visual inspection of the Bland–Altman-plots (Supplementary Figure 2) did not reveal any systematic deviations in the sense of heteroscedasticity, particularly no indication that variability between test–retest is higher in more pathological EPs.

### *Variance components from mixed linear effect models*

Models were calculated on the measures with highest concordance. Total absolute variance and variance components are given in Tables 3 and 4. Subject-related variance components explained most of the overall variability in single EP modalities (SEP-T: 73%; SEP-M: 81%; MEP-UL shortest CxM: 89%, and >90% in the remainder) and in quantitative EP scores (>90%, except qSEP: 76%). Rater-related variance (including the error term) explained the following proportion of variability: MEPs up to 5.5%, SEP-M: 10%, SEP-T: 16%; mqEPS-short: 6.9%, mqEPS-mean: 5.8%, qMEP-short: 3.9%, qMEP-mean: 2.2%, qSEP: 12%.

### *Measurement-related variability and critical difference D*

Tables 3 and 4 give the SE for single EP modalities and EP scores for a comparison of two equal sized groups with 20 subjects per arm cross-sectionally ($SE_{cross}$), in regard to their longitudinal change ($SE_{long}$) and for a within-subject longitudinal comparison ($SE_{1S}$; $n$=1) with respective critical differences ($D_{MDC}$, $D_{imp}$). For MEPs, mqEP and qMEP two values are given: the first refers to scores including the shortest CxM, the second to scores including mean CxM.

With 20 subjects per arm, a cross-sectional group difference ($D_{cross}$) of 2.3 and 2.4 ms in MEP-UL, 3.7 and 3.9 ms in MEP-LL, 0.7 ms in SEP-M and 3.3 ms in SEP-T, and group difference in longitudinal changes ($D_{long}$) of 0.9 and 0.8 ms in MEP-UL, 1.1 and 0.8 ms in MEP-LL, 0.4 ms in SEP-M and

**Table 3.** Variance components from linear mixed effect models for single EP modalities.

| (a) | MEP-UL | | MEP-LL | | SEP-M | SEP-T |
|---|---|---|---|---|---|---|
| | Shortest CxM | Mean CxM | Shortest CxM | mean CxM | N20 | P40 |
| Total Var (ms) | 13.80 | 14.89 | 36.37 | 39.30 | 1.30 | 31.44 |
| S | 9.75 | 11.47 | 29.79 | 33.42 | 0.89 | 22.73 |
| R | 0.06 | 0.12 | 0.13 | 0.19 | 0.01 | 0.25 |
| S*T | 0.03 | 0.09 | 0.34 | 0.24 | 0.09 | 3.25 |
| S*R | 0.01 | 0.00 | 0.00 | 0.02 | 0.00 | 0.74 |
| S*L | 2.60 | 2.35 | 3.88 | 4.43 | 0.16 | 0.16 |
| R*T | 0.05 | 0.00 | 0.12 | 0.02 | 0.01 | 0.01 |
| S*R*T | 0.15 | 0.10 | 0.34 | 0.15 | 0.07 | 0.82 |
| S*L*R | 0.00 | 0.00 | 0.14 | 0.05 | 0.00 | 2.38 |
| S*L*T | 0.67 | 0.55 | 0.64 | 0.38 | 0.02 | 0.31 |
| Error | 0.50 | 0.20 | 0.99 | 0.41 | 0.04 | 0.81 |
| **$SE_{cross}$** | **1.16** | **0.85** | **1.33** | **1.39** | **0.25** | **1.20** |
| **$SE_{long}$** | **0.45** | **0.40** | **0.57** | **0.42** | **0.19** | **0.93** |
| **$SE_{1S}$** | **1.45** | **1.25** | **1.85** | **1.35** | **0.59** | **2.96** |
| **$D_{MDC}$** | **2.83** | **2.46** | **3.62** | **2.64** | **1.16** | **5.80** |
| **$D_{imp}$** | **1.85** | **1.60** | **2.36** | **1.72** | **0.76** | **3.79** |

Total variance and single variance components including interaction terms are given in the upper part (Var: variance; S: subject; R: rater; T: time; L: side).
UL: upper limb; LL: lower limb; SEP-M: median SEP; SEP-T: tibial SEP; CxM: cortico-muscular latency.
Related standard errors ($SE_{cross}$, $SE_{long}$, $SE_{1S}$) and critical differences ($D_{MDC}$, $D_{imp}$) are given in bold, please refer to the main text for details.

**Table 4.** Variance components from linear mixed effect models for quantitative EP scores

| | Multimodal qEPS | | Unimodal qEPS | | |
|---|---|---|---|---|---|
| | mqEPS-short | mqEPS-mean | qMEP-short | qMEP-mean | qSEP |
| Total Var (ms) | 3.63 | 3.91 | 6.39 | 7.34 | 3.17 |
| S | 3.27 | 3.58 | 5.93 | 7.05 | 2.42 |
| R | 0.03 | 0.04 | 0.05 | 0.08 | 0.03 |
| S*T | 0.11 | 0.11 | 0.21 | 0.14 | 0.37 |
| S*R | 0.03 | 0.06 | 0.00 | 0.00 | 0.09 |
| R*T | 0.01 | 0.00 | 0.03 | 0.00 | 0.01 |
| S*R*T | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Error | 0.18 | 0.13 | 0.17 | 0.08 | 0.25 |
| **$SE_{cross}$** | **0.69** | **0.75** | **0.79** | **0.85** | **0.54** |
| **$SE_{long}$** | **0.26** | **0.24** | **0.24** | **0.19** | **0.32** |
| **$SE_{1S}$** | **0.83** | **0.76** | **0.78** | **0.60** | **1.00** |
| **$D_{MDC}$** | **1.62** | **1.49** | **1.54** | **1.18** | **1.96** |
| **$D_{imp}$** | **1.06** | **0.98** | **1.00** | **0.77** | **1.28** |

Total variance and single variance components including interaction terms are given in the upper part (Var: variance; S: subject; R: rater; T: time; mqEPS: modified quantitative EP score; qMEP: quantitative MEP-score; qSEP: quantitative SEP score).
Related standard errors ($SE_{cross}$, $SE_{long}$, $SE_{1S}$) and critical differences ($D_{MDC}$ and $D_{imp}$) are given in bold, please refer to the main text for details.

1.8 ms in SEP-T reflects a statistically significant change at an alpha level of 5%. Values in EP scores are lower: 1.2 and 1.2 in mqEPS, 1.6 and 1.7 in qMEP and 1.1 in qSEP for $D_{cross}$, and 0.39 and 0.36 in mqEPS, 0.48 and 0.37 in qMEP and 0.62 in qSEP for $D_{long}$. These values decrease by 50% if the group size is quadrupled.

MDC is relatively high with values of 2.8 and 2.5 ms in MEP-UL, 3.6 and 2.6 ms in MEP-LL, 1.2 ms in SEP-M and 5.8 ms in SEP-T, MDC in EP scores is lower: 1.6 and 1.5 in mqEPS, 1.5 and 1.2 in qMEP and 2.0 in qSEP.

For a more sensitive detection of improvement or progression, a one-sided testing at an alpha level of 10% yields lower critical differences. $D_{imp}$ are 1.9 and 1.6 ms in MEP-UL, 2.4 and 1.7 ms in MEP-LL, 0.8 ms in SEP-M and 3.8 ms in SEP-T. For EP scores $D_{imp}$ are 1.1 and 1.0 in mqEPS, 1.0 and 0.8 in qMEP and 1.3 in qSEP.

## Discussion

Based on a sample of 21 MS patients with relapsing and progressive disease course we investigated the measurement variability of SEPs, MEPs and associated quantitative EP scores with a median test–retest interval of 8 days using a standardised recording protocol in three centres. As pathological changes during such a period are unlikely, the within-subject test–retest differences define the measurement error. To account for physiological and rater variability, each curve was independently assessed by each of the five raters using a custom-made server-based software (EPMark), and retest reading was done blinded to baseline results.

First-level analysis showed that the main cortical responses in SEPs and the cortico-muscular latencies in MEPs have lower measurement error than the central sensory and motor conduction times, which has been shown previously for MEPs.[16] This result is most likely due to the fact that central conduction is determined as the difference of overall and peripheral conduction in which the latter introduces additional variability into the measurement. The distinction between central and peripheral conduction yields important diagnostic information. However, in the context of a longitudinal within-subject assessment in patients with MS the situation is different. Here, changes in the overall conduction are attributable to the central part as peripheral conduction can be assumed to remain stable except in patients with concomitant diseases such as myelopathy not due to MS, diabetic and other polyneuropathies among others. However, these patients are usually excluded from participation in clinical trials.

Due to pre-innervation for facilitation, MEPs can exhibit baseline fluctuations and the rater has to make some estimations regarding the onset. Physiologically, fluctuations in cortical and spinal excitability influence the spatial and temporal summation of incoming volleys at the spinal motoneuron.[27] Experimentally, the mean within-session trial-to-trial difference has been shown to amount to 0.59 ms (SD 0.17) in healthy controls and to 1.49 ms (SD 1.23) in patients with MS and the authors have even proposed to use this variability as a diagnostic sign.[21] Therefore, the mean instead of the shortest cortico-muscular latency has advantages for longitudinal within-subject comparisons, and averaging has been shown to increase reliability.[29] Rater-related variability is potentially reduced by consensus reading of discrepant curves and by rating with comparison to previous curves. To account for rater-related variability, we deliberately renounced from these procedures. However, comparison of successive curves is especially important in SEPs in which the higher measurement error in our study is in large part due to single outlying subjects with ambiguous peaks.

The yield of pathological tests is important for measuring a therapeutic response on slowed conduction. The chance of a pathologically prolonged latency correlates with the length of the tract as also shown in our data, in which the rate of pathological results was slightly lower compared to rates reported in the literature in more advanced patients.[6,30]

At second-level analysis, we determined the measurement variability and estimated the critical difference for group and single subject settings assuming a central reading by two independent raters.

In single EP modalities, the critical difference for a cross-sectional comparison of 20 subjects per arm lies around 2 ms for cross-sectional and around 1 ms for comparison of longitudinal changes, and is numerically lower in quantitative EP scores. Achieving such group differences seems to be likely given the results of two recent VEP trials.[31,32] They have demonstrated a mean difference in longitudinal changes by 6.1 ms at 32 weeks in favour of patients treated with opicinumab and a mean within-subject shortening of the VEP latency by 1.7 ms under treatment with clemastine. In an observational study, a quantitative EP score increased from 2.6 to 3.5 in a sample of 72 relapsing and progressive MS patients after one year.[24]

At the single subject level, the context of use and the translation into clinical relevance has to be considered when deciding on the false positive rate. The minimal detectable change (two-sided, alpha level: 5%)[14] is numerically quite high and a patient with a change in a single modality of more than 3 ms or a change in a quantitative EP score of more than 1.5 is very likely also to have signs of clinical progression (Figure 3 in Hardmeier et al.).[11] In clinical practice, sensitive measures are required to identify patients before they progress, and in clinical studies, which compare responder rates between treatment arms as the primary outcome, a higher event rate may be preferable at the cost of some specificity. In both settings, a one-sided testing would be appropriate and a false positive rate of 10% acceptable, resulting in a critical difference in the range of 1.6–2.4 ms for single EP modalities (except SEP-T: 3.8 ms) and in the range of 0.8–1.3 for quantitative EP scores. Depending on the interval between the assessments and the effects of an intervention, these numbers seem to be realistic. More empirical data have to be gathered to determine the clinical relevance of these cut-offs. However, a 6-month change in a quantitative EP score is predictive of an EDSS progression at 3 years at the group level.[24]

EP scores may have several advantages over single EP modalities. Their validity is higher as they are more closely associated with overall clinical disability. As the mean of several tests, they are less influenced by single outlying tests and less prone to selection bias. The latter may occur if only the pathological tests of a multimodal assessment are taken at baseline to measure treatment effects at follow-up as a regression to the mean may simulate improvement.[33] However, systematic effects related to the state of the subject may sum up in a score. In stable patients, fluctuations of cortical and spinal excitability seem to play on a time scale of seconds rather than days,[21] as the reported intra-session variability in MEPs is quite comparable to the inter-session variability determined in our study. In contrast, disease activity may influence cortical excitability,[34] and a conduction block in an affected tract clearly changes the EP response. No measurable effect on EP latencies has been demonstrated in fatigue making it unlikely to be a confounding factor,[35,36] while sodium blocking agents and 3-4-aminopyridine probably are. In our study, EPs were recorded at least 80 days after a last relapse and we controlled for change in potentially confounding medication. Hence, we consider the absolute

reliability determined in our sample a realistic estimate for studies in MS.

Some limitations have to be acknowledged. While magnetic and electric stimuli are well standardised and robust to hardware and software updates, recording of EPs is time consuming and accuracy depends on the strict adherence to the recording protocol. The reading of EP curves still needs experienced neurophysiologists, who, however, can mark curves on a standardised display totally independent from each other and blinded to clinical information by using EPMark. While test–retest studies frequently have only samples of less than 20 subjects,[13] a larger sample would have given closer estimates of the variance components in our model. However, as these components clearly differ from each other qualitatively, different conclusions from our data are quite unlikely. The responsiveness of EP scores to therapeutic interventions are currently not well known whereas sensitivity to change has been shown several times in groups of patients.[11]

## Conclusions

In summary, standardised multicentre EP assessment with central reading is feasible and reliable. Mean cortico-muscular latency in MEPs and the main cortical responses in SEPs have higher reliability compared with central conduction times. EP scores are less influenced by outlying tests and are more closely related to overall clinical disability. A comparison of longitudinal changes between two groups has a smaller critical difference than a cross-sectional comparison. In both settings, significant group changes seem realistically achievable in small samples. At a single subject level, the cut-offs defining improvement or progression with sufficient sensitivity remain to be determined. The results support the concept of using EPs as a candidate response biomarker in clinical trials in MS for quantification of disease progression and for studying remyelination. As our tool for central reading of multimodal EPs (EPMark) is operational, larger multicentre trials are warranted to corroborate the current results and to determine the mean longitudinal change in EP scores in well-defined patient cohorts for precise sample size estimation.

## Acknowledgements

The authors would like to thank all the patients and the technical staff for their participation.

## Conflict of Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: MH's institution has received fees from Roche for consultancy services; his research is or was supported by the Swiss Multiple Sclerosis Society and the Swiss National Science Foundation SPUM 33CM30_124115 and 33CM30_140338.
FJ has received honorariums and unrestricted grants from Biogen, Sanofi-Genzyme, Merck-Serono for participating in ad boards, doing presentations and for conducting investigator-initiated trials.
PA reports grants, personal fees and non-financial support from Allergan, Biogen, Ipsen, Merz Pharmaceuticals, Novartis, and Roche, personal fees and non-financial support from Bayer Healthcare and Merck, and non-financial support from Sanofi-Aventis/Genzyme.
HB has no conflicts of interest to declare.
CS has no conflicts of interest to disclose.
LL received honoraria for consulting services from Merck, Roche, Biogen and for speaking activities from Teva; research support from Merck, Biogen, Novartis; travel support from Merck, Roche, Biogen, Almirall.
PF's research is or was supported by the Swiss National Science Foundation SPUM 33CM30_124115 and 33CM30_140338 (PI), Swiss Multiple Sclerosis Society, Synapsis Foundation, Parkinson Schweiz, Novartis Research Foundation, Gossweiler Foundation, Freiwillige Akademische Gesellschaft Basel, Mach-Gaensslen-Stiftung, Botnar Foundation, Bangerter Foundation, and by unconditional research grants from industry (Roche, AbbVie, Biogen, General Electrics).

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Study funding was provided by an unrestricted, investigator-initiated trial grant from Biogen Inc. MA, USA.

## ORCID iD

Martin Hardmeier http://orcid.org/0000-0003-1451-7337
Letizia Leocani http://orcid.org/0000-0001-9326-6753

## Supplemental Material

Supplemental material for this article is available online.

## References

1. Amur S, LaVange L, Zineh I, et al. Biomarker qualification: toward a multiple stakeholder framework for biomarker development, regulatory acceptance, and utilization. *Clin Pharmacol Ther*. 2015; 98: 34–46.
2. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of

measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010; 63: 737e45.
3. McDonald WI. Pathophysiology of conduction in central nerve fibres. In: Desmedt JE (ed) *Visual Evoked Potentials in Man: New Developments*. Oxford, UK: Clarendon Press, 1977, pp. 427–437.
4. Smith KJ. Conduction properties of central demyelinated and remyelinated axons, and their relation to symptom production in demyelinating disorders. *Eye (Lond)* 1994; 8: 224–237.
5. Fuhr P, Borggrefe-Chappuis A, Schindler C,et al. Visual and motor evoked potentials in the course of multiple sclerosis. *Brain* 2001; 124: 2162–2168.
6. Leocani L, Rovaris M, Boneschi FM, et al. Multimodal evoked potentials to assess the evolution of multiple sclerosis: a longitudinal study. *J Neurol Neurosurg Psychiatry* 2006; 77: 1030–1035.
7. Schlaeger R, D'Souza M, Schindler C, et al. Combined evoked potentials as markers and predictors of disability in early multiple sclerosis. *Clin Neurophysiol* 2012; 123: 406–410.
8. Schlaeger R, D'Souza M, Schindler C, et al. Electrophysiological markers and predictors of the disease course in primary progressive multiple sclerosis. *Mult Scler* 2014; 20: 51–56.
9. Schlaeger R, Schindler C, Grize L, et al. Combined evoked potentials predict MS disability after 20 years. *Mult Scler* 2014; 20: 1348–1354.
10. London F, Sankari SE and van Pesch V. Early disturbances in multimodal evoked potentials as a prognostic factor for long-term disability in relapsing–remitting multiple sclerosis patients. *Clin Neurophysiol*. 2017; 128: 561–569.
11. Hardmeier M, Leocani L and Fuhr P. A new role for evoked potentials in MS? Repurposing evoked potentials as biomarkers for clinical trials in MS. *Mult Scler* 2017; 23: 1309–1319.
12. Atkinson G and Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998; 26: 217e38.
13. Beaulieu LD and Milot MH. Changes in transcranial magnetic stimulation outcome measures in response to upper-limb physical training in stroke: a systematic review of randomized controlled trials. *Ann Phys Rehabil Med* 2018; 61: 224–234.
14. Beckerman H, Roebroeck ME, Lankhorst GJ,et al. Smallest real difference, a link between reproducibility and responsiveness. *Qual Life Res* 2001; 10: 571e8.
15. Schambra HM, Ogden RT, Martinez-Hernandez IE, et al. The reliability of repeated TMS measures in older adults and in patients with subacute and chronic stroke. *Front Cell Neurosci* 2015; 9: 335.
16. Livingston SC and Ingersoll CD. Intra-rater reliability of a transcranial magnetic stimulation technique to obtain motor evoked potentials. *Int J Neurosci* 2008; 118: 239e56.
17. Hoonhorst MH, Kollen BJ, van den Berg PS, et al. How reproducible are transcranial magnetic stimulation induced motor evoked potentials in subacute stroke? *J Clin Neurophysiol* 2014; 31: 556e62.
18. Cacchio A, Cimini N, Alosi P, et al. Reliability of transcranial magnetic stimulation-related measurements of tibialis anterior muscle in healthy subjects. *Clin Neurophysiol* 2009; 120: 414e9.
19. Brown KE, Lohse KR, Mayer IMS, et al. The reliability of commonly used electrophysiology measures. *Brain Stimul* 2017; 10: 1102–1111.
20. Troni W, Melillo F, Bertolotto A, et al. Normative values for intertrial variability of motor responses to nerve root and transcranial stimulation: a condition for follow-up studies in individual subjects. *PLoS One* 2016; 11: e0155268.
21. Britton TC, Meyer BU and Benecke R. Variability of cortically evoked motor responses in multiple sclerosis. *Electroencephalogr Clin Neurophysiol* 1991; 81: 186–194.
22. Matthews WB and Small DG. Serial recording of visual and somatosensory evoked potentials in multiple sclerosis. *J Neurol Sci* 1979; 40: 11–21.
23. Vogel P and Vogel H. Somatosensory cortical potentials evoked by stimulation of leg nerves: analysis of normal values and variability; diagnostic significance. *J Neurol* 1982; 228: 97–111.
24. Schlaeger R, Hardmeier M, D'Souza M, et al. Monitoring multiple sclerosis by multimodal evoked potentials: numerically versus ordinally scaled scoring systems. *Clin Neurophysiol* 2016; 127: 1864–1871.
25. Mauguière F, Allison T, Babiloni C, et al. Somatosensory evoked potentials. The International Federation of Clinical Neurophysiology. *Electroencephalogr Clin Neurophysiol Suppl* 1999; 52: 79–90.
26. Rothwell JC, Hallett M, Berardelli A, et al. Magnetic stimulation: motor evoked potentials. The International Federation of Clinical Neurophysiology. *Electroencephalogr Clin Neurophysiol Suppl* 1999; 52: 97–103.
27. Rossini PM, Burke D, Chen R, et al. Noninvasive electrical and magnetic stimulation of the brain, spinal cord, roots and peripheral nerves: basic principles and procedures for routine clinical and research application. An updated report from a IFCN Committee. *Clin Neurophysiol* 2015; 126: 1071e107.
28. Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1: 307–310.
29. Bastani A and Jaberzadeh S. A higher number of TMS-elicited MEP from a combined hotspot improves intra- and inter-session reliability of the upper limb muscles in healthy individuals. *PLoS One* 2012; 7: e47582.
30. Chiappa KH. *Evoked potentials in clinical medicine*, 3rd ed. Philadelphia, PA: Lippincott-Raven, 1997. Chapters 8 (6.1, 6.2, 9.1.6) and 9 (1.1.3, 3.2).

31. Cadavid D, Balcer L, Galetta S, et al.; RENEW Study Investigators. Safety and efficacy of opicinumab in acute optic neuritis (RENEW): a randomised, placebo-controlled, phase 2 trial. *Lancet Neurol* 2017; 16: 189–199.
32. Green AJ, Gelfand JM, Cree BA, et al. Clemastine fumarate as a remyelinating therapy for multiple sclerosis (ReBUILD): a randomised, controlled, double-blind, crossover trial. *Lancet* 2017; 390: 2481–2489.
33. Barnett AG, van der Pols JC and Dobson AJ. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol* 2005; 34: 215–220.
34. Caramia MD, Palmieri MG, Desiato MT, et al. Brain excitability changes in the relapsing and remitting phases of multiple sclerosis: a study with transcranial magnetic stimulation. *Clin Neurophysiol* 2004; 115: 956–965.
35. Liepert J, Mingers D, Heesen C, et al. Motor cortex excitability and fatigue in multiple sclerosis: a transcranial magnetic stimulation study. *Mult Scler* 2005; 11: 316–321.
36. Schlaeger R, Hardmeier M and Fuhr P. Superficial brain stimulation in multiple sclerosis. *Handb Clin Neurol* 2013; 116: 577–584.