Monitoring Multiple Sclerosis by Multimodal Evoked Potentials: numerically versus ordinally scaled scoring systems

Regina Schlaeger, MD¹; Martin Hardmeier, MD¹; Marcus D'Souza, MD¹; Leticia Grize, PhD^{2,3}; Christian Schindler, PhD^{2,3}; Ludwig Kappos, MD^{1,3}; Peter Fuhr, MD^{1,3}

Department of Neurology, University Hospital Basel (1), Swiss Tropical and Public Health Institute (2), University of Basel (3), Switzerland

Corresponding author:

Prof. Dr. Peter Fuhr

University Hospital Basel, Department of Neurology, Petersgraben 4, 4031 Basel, Switzerland

Tel.: +41-612654167

Fax: +41-612654100

E-mail: peter.fuhr@usb.ch

Keywords: Multiple sclerosis, prognosis, monitoring, evoked potentials.

Abstract

Objective: To compare the ability of different evoked potential scores (EPS) to monitor and predict the disease course in multiple sclerosis (MS).

Methods: Seventy-two patients with MS or clinically isolated syndrome were investigated by visual, motor, and somatosensory EP and Expanded Disability Status Scale (EDSS) at baseline (T0) and months 6, 12, 24, 36 (T4). EP results were rated according to ordinal (o), semi-quantitative (sq), and quantitative (q) EPS. Spearman rank correlation and multivariable linear regression were used to investigate the associations between EPS and clinical disability.

Results: All EPS correlated with EDSS cross-sectionally ($0.72 \leq \rho \leq 0.87$, all $p < 0.001$) and longitudinally ($0.39 \leq \rho \leq 0.47$, all $p \leq 0.004$). EPS_{T0} and EDSS_{T0} together explained 85-86% of EDSS_{T4} variance. A posteriori power calculation showed that the sample sizes needed to detect significant changes over 6 months in q-EPS, sq-EPS and o-EPS with 90% certainty would be 50, 129 and 222, respectively. q-EPS change_{T1-T0} correlated with EDSS change_{T4-T0} ($\rho = 0.56$, $p < 0.001$), while sq-EPS and o-EPS changes_{T1-T0} did not.

Conclusion: All three EPS allow disease course monitoring in MS. However, the quantitative EPS detects clinically relevant short-term changes with a smaller sample size than semi-quantitative or ordinal EPS.

Significance: These results underscore the potential of EPS to characterize MS disease evolution.

Introduction

Multimodal evoked potentials (EP) allow monitoring and prediction of MS disability in small groups of patients as was shown by our group and several others (Fuhr et al., 2001; Leocani et al., 2006; Kallmann et al., 2006; Jung et al., 2008; Invernizzi et al., 2011; Margaritella et al., 2012a; Schlaeger et al., 2012a, b, 2014a, b).

To summarize multimodal EP by a one-dimensional statistical measure, different scoring systems have been applied. Some groups used *ordinal* scores based on qualitative assessments of EP latencies, and amplitudes (O'Connor et al., 1998; Kallmann et al., 2006) or absence of principal EP components (Leocani et al., 2006, Invernizzi et al., 2011). Others applied more complex *semi-quantitative* scores with a step-wise rating of the extent of latency prolongation in relation to the upper limit of normal (Jung et al., 2008; Margaritella et al., 2012a) or *quantitative* scores based on z-transformed latencies (Fuhr et al., 2001, Schlaeger et al., 2012a and 2014a).

However, there is currently no consensus on how to best score EP.

The goal of this study is to compare different EP scores (EPS) and their ability to monitor and predict the disease course in a cohort of patients with MS. For this purpose, we applied an ordinal and a semi-quantitative EPS on our previously quantitatively analyzed cohorts of relapsing-remitting and primary progressive MS patients (Schlaeger et al. 2012a, Schlaeger et al. 2014a).

Methods

Patients

The current analysis is based on the data sets of two previously reported MS cohorts that were described in (Schlaeger et al. 2012a, and 2014a). Inclusion criteria comprised either a diagnosis of MS according to the McDonald criteria (McDonald et al., 2001) or a diagnosis of a clinically isolated syndrome (CIS). Patients with primary progressive MS additionally fulfilled the Thompson criteria (Thompson et al., 2000). Exclusion criteria were contraindications for magnetic evoked potentials, acute relapses or glucocorticosteroid treatment within 6 weeks prior to baseline examination, additional neurological diseases potentially interfering with MS symptoms, or severe additional other diseases.

Investigations consisted of a standardized neurological examination including assessment of the EDSS / Neurostatus (Kurtzke, 1983; Kappos et al., 2007), and multimodal EP. For the current analysis the examinations at entry (T0), months 6 (T1), 12 (T2), 24 (T3), and 36 (T4) were selected. If a patient had a relapse at the assigned study visit, the clinical and EP examinations were postponed to at least 6 weeks after the first relapse symptoms. Patients who decided to discontinue the study before T4 were again invited to our center for a neurological examination or underwent a structured standardized telephone interview for determination of the EDSS by phone (Lechner-Scott et al., 2003) at T3 and T4.

The local ethic committee of the University of Basel approved the study. Written informed consent was obtained from all participants.

All seventy-two participants of the two cohorts were included into this study: 50% women, mean age: 42 years (SD=11.4), mean disease duration: 6.6 years (SD=4.75), median EDSS at baseline: 2.5 (IQR: 1.5-3.5), median EDSS at year 3:

3.0 (IQR: 2.0-4.0). Forty-seven patients had relapsing (RR) MS, three a CIS and twenty-two primary progressive (PP) MS. At baseline, 44% of patients were treated with immunomodulatory drugs (glatirameracetate, interferon-b), 1.3% received immunosuppressive treatments (mitoxantrone).

A selective dropout of patients with a more severe disease course reduced the dispersion of the EP data at later time-points with an apparent ceiling effect of EPS. We were able to obtain the EDSS or EDSS by phone at T4 of fourteen out of nineteen dropouts, so that the clinical data of 93% of all participants was available for the predictive analysis. Patients with complete clinical and electrophysiological data sets at T4 had a median EDSS_{T4} of 3.0 (IQR 2-4)), while patients who dropped out of the electrophysiological part of the study had a median EDSS_{T4} of 4.5 (IQR 2.125-6.375).

Evoked potentials

Recording of visual evoked potentials (VEP), motor evoked potentials (MEP) to upper and lower extremities (UE/LE) and median and tibial nerve somatosensory evoked potentials (SEP) followed closely the recommendations of the International Federation of Clinical Neurophysiology (Mauguière et al. 1999, Celesia et al. 1999, Rothwell et al. 1999) as described previously (Schlaeger et al., 2012a).

Briefly, full-field VEP were recorded from an active electrode placed 3cm above Oz and a reference electrode at Fz. Pattern reversal stimulation was presented to each eye separately at a frequency of 0.5Hz. The distance from the eye to the screen was 225 cm, each checkerboard field was 30 minutes of arc. VEP's were averaged twice, 256 times.

MEP were recorded from the abductor digiti minimi (ADM) and the tibial anterior (TA) muscles bilaterally. The muscles were pre-innervated by the participants. The

examiner controlled the extent of contraction both manually and acoustically.

Magnetic stimuli were delivered first to the hand and secondly to the leg areas of the motor cortex with a Magstim 200 device (The Magstim Company Ltd, Whitland, UK) via a round coil using maximal output of the stimulator (2.2T). Four stimulations were done with the coil pointing in clock-wise current direction and four stimulations were done with the coil pointing in counter-clockwise current direction, for ADM and TA tests, respectively. The fastest conduction time out of these stimulations was used to calculate the central motor conduction time (CMCT). For spinal stimulation the rim of the coil was placed over the seventh cervical and fifth lumbar vertebra. Median and tibial nerve SEP were obtained by stimulating the median nerve at the wrist and the posterior tibial nerve at the ankle, respectively. The stimulus intensity exceeded visual motor threshold by about 3mA. The stimulus frequency was 3Hz. For the median nerve SEP active recording electrodes were placed subcutaneously between cervical vertebra 6 and 7 and 7cm lateral and 2cm posterior to Cz with the reference electrode placed at Fz. For the tibial nerve SEP active recording electrodes were placed between the 12th thoracic and first lumbar vertebra and 3cm posterior to Cz with the reference electrodes placed subcutaneously near the anterior superior iliac spine and Fz, respectively.

Normal values were taken from our own laboratory for VEP and tibial nerve SEP (height adjusted) and from Stoehr et al., 1996 for median nerve SEP and MEP (height adjusted). The P100 latency of the VEP, the latency difference N13-N20 of the median nerve SEP, the P40 latency of the tibial SEP and the CMCT to upper and lower extremities were analyzed and regarded as pathological when exceeding the upper limit of normal (mean + 2.5 SD).

A VEP amplitude was regarded as pathologically reduced if either the P100-N145 amplitude was < 5µV or if the inter-side difference exceeded > 50% corresponding to

an amplitude ratio > 2 . In case of a W-shaped morphology VEP amplitudes were regarded as pathological if the difference between the two positive peaks exceeded 10ms (Kallmann et al., 2006).

MEP amplitudes were regarded as pathologically reduced if the amplitude ratio between the MEP and the corresponding compound motor action potential (to stimulation of the peripheral nerve) was $< 17\%$ in upper extremities and $< 10\%$ in lower extremities (according to Stoehr et al., 1996) or in case of an inter-side difference $> 50\%$ according to available definitions (Kallmann et al., 2006, Jung et al., 2008).

SEP amplitudes were considered reduced if the amplitude N20-P25 was $< 0.9\mu\text{V}$ in median SEP (according to Stoehr et al, 1996) or if the amplitude N35-P40 was $< 0.3\mu\text{V}$ in tibial SEP (according to Kallmann et al., 2006 who used the same SEP methodology as applied in our lab).

As judgments of amplitude morphology imply a subjective element, two experienced clinical neurophysiologists (RS, MH) independently rated the ordinal EPS (o-EPS) that includes amplitude morphology as a criterion in addition to latency. In case of discordant results a consensus reading was performed. The EP raters were blinded to the patients' EDSS. The inter-rater reliability for the o-EPS was good with a Cohen's kappa of 0.77.

EP Scores (EPS)

Supplementary Figure S1 illustrates the EP evaluation according to the three following scoring systems based on the EP raw data of a patient:

- (A) *Ordinal EPS (o-EPS)*: In analogy to Leocani et al., 2006, EP results were scored as follows: 0 = normal; 1 = increased latency; 2 = increased latency plus morphological abnormalities of a major component; 3 = absence of a

major component; maximum score = 30 points (3 points x 2 sides x 5 EP modalities).

(B) *Semi-quantitative* EPS (sq-EPS): According to Jung et al., 2008, EP results were scored as: 0 = normal; 1 = pathological side difference of latency; 2 = latency above the normal range but below $1.1 \times$ upper limit, or $> 50\%$ side difference of amplitude; 3 = latency $1.1\text{--}1.3 \times$ upper limit; 4 = latency above $1.3 \times$ upper limit; 5 = absent EP component. If both the latency and amplitude of a major EP component was abnormal, the result producing the higher single EP score was taken into account. Thus, the worst possible result was 50 points (5 points x 2 sides x 5 EP modalities).

(C) *Quantitative* EPS (q-EPS): The sum of z-transformed EP latencies of the right and left side of P100, CMCT_{UE}, CMCT_{LE}, N13 – N20 and P40 was calculated (Schlaeger et al., 2014a). We tested two imputation methods for absent cortical components of MEP, SEP or VEP: a) substitution by the longest central conduction time observed in the whole cohort and within the same EP modality (Fuhr et al., 2001), and b) by the last measured latency of the individual. The maximal z-values for each EP modality were as follows: z-P100 = 19.6; z-CMCT_{UE} = 22.6; z-CMCT_{LE} = 30.06, z-N13-N20 = 30.41; z-P40 = 32.53.

The scores A-C were divided by the number of EP modalities performed, in order to not lose data of those patients who were not able to cooperate in an examination of single EP modalities (e.g. due to leg cramps or back pain in tibial SEP) at single points in time. The mean and median number of modalities examined at all time points was 9.9 and 10 out of 10.

Statistical analysis

To assess short-term changes within 6 and 12 months for each EPS (ΔEPS_{T1-T0} , ΔEPS_{T2-T0}) and to test the global null hypothesis of no average change between time points 0, 6 and 12 months, a mixed linear model with a random intercept for patient and a fixed effect for time was used. Moreover, assuming that the observed changes coincided with the true average changes at population level, we calculated the sample size needed to detect a significant EPS change over 6 and 12 months, respectively, with a power of 0.90 at the usual significance level of 0.05 using a 2-tailed test.

Spearman rank correlation was used to examine 1) the *cross-sectional* associations between each EPS and the EDSS at each time point (T0-T4), 2) the *longitudinal* associations between change in each EPS over 3 years (ΔEPS_{T4-T0}) with the corresponding change in EDSS ($\Delta\text{EDSS}_{T4-T0}$), and 3) the *predictive* associations between each EPS at baseline (EPS_{T0}), its change over 6 months (ΔEPS_{T1-T0}) with the two measures of clinical outcome, EDSS_{T4} and $\Delta\text{EDSS}_{T4-T0}$. Bootstrapping was performed with 1000 iterations to estimate and compare the confidence intervals of the correlation coefficients obtained for the respective scores.

Moreover, we performed a multivariable regression analysis to examine the predictive relationship of each EPS_{T0} and its corresponding short-term change ΔEPS_{T1-T0} with EDSS_{T4} as outcome variable, adjusting for other potentially relevant baseline variables (sex, age, EDSS_{T0}). The starting models are referred to as Model 1a-c, and contain both EPS_{T0} and ΔEPS_{T1-T0} along with all other baseline variables. Predictor variables with p-values >0.20 (sex, age, ΔEPS_{T1-T0}) were eliminated from these models in a backward selection process resulting in Models 2a-c. Model comparisons were based on the amount of explained variability adjusted for number of predictors (adjusted R^2) and the Bayesian information criterion (BIC). The BIC

reduces the risk of over-fitting by penalizing the complexity of the model.

Receiver operating characteristic (ROC) curves were determined to compare the prediction of clinical worsening over 3 years given by binary logistic models based on 1) the baseline variables EPS_{T0} as single predictors, 2) the variables of short-term EPS change ΔEPS_{T1-T0} and 3) the combinations of EPS_{T0} and ΔEPS_{T1-T0} of each scoring system, respectively. Worsening thereby was defined as an increase in EDSS by 1.5 points or more if the baseline EDSS was 0, by 1.0 point or more if the baseline EDSS was between 1 and 5.0 and by 0.5 points or more if the baseline EDSS was > 5.0 (Calabresi et al., 2014). We calculated, the areas under the ROC curve for the different models and compared them pairwise between the different EPS by using bootstrapping. Re-sampling was done 1000 times.

Statistical analysis was done with SPSS version 14.0 (Chicago, IL: SPSS Inc. 2005) and SAS version 9.3 (Cary, N.C.: SAS Institute 2002-2010).

Results

EPS and short-term change in EPS

The distribution of the EP raw data is summarized in the Supplementary Table S1, the distribution of the EPS values at each time point is given in Table 1.

While EP changes measured on the quantitative and semi-quantitative scales reached statistical significance after 6 months ($\Delta q\text{-EPS}_{T1-T0}$: $p < 0.001$; $\Delta q\text{-EPS}_{T2-T0}$: $p < 0.001$; $\Delta sq\text{-EPS}_{T1-T0}$: $p = 0.02$; $\Delta sq\text{-EPS}_{T2-T0}$: $p = 0.001$), EP changes measured on the ordinal scale were in the same direction but did not reach statistical significance ($\Delta o\text{-EPS}_{T1-T0}$: $p = 0.13$; $\Delta o\text{-EPS}_{T2-T0}$: $p = 0.15$). On the quantitative scale statistically significant differences ($p < 0.001$) were observed regardless of the imputation proceeding used for absent cortical components.

In a sensitivity analysis we only considered full data sets for model calculation. While EP changes measured on the quantitative and semi-quantitative scales reached statistical significance after 6 months ($\Delta q\text{-EPS}_{T1-T0}$: $p < 0.001$; $\Delta q\text{-EPS}_{T2-T0}$: $p < 0.001$; $\Delta sq\text{-EPS}_{T1-T0}$: $p = 0.048$; $\Delta sq\text{-EPS}_{T2-T0}$: $p = 0.004$), EP changes measured on the ordinal scale were in the same direction but did not reach statistical significance ($\Delta o\text{-EPS}_{T1-T0}$: $p = 0.24$; $\Delta o\text{-EPS}_{T2-T0}$: $p = 0.19$).

If our observed changes were the true average changes at population level, the sample size needed to detect a significant EPS change over 6 months (12 months) with a power of 0.9 would be 50 (34) for the q-EPS, 129 (55) for the sq-EPS and 222 (206) for the o-EPS.

Cross-sectional and longitudinal correlations between EPS and clinical outcome

All three EP-scores similarly correlated with the EDSS in cross-section ($0.72 \leq \rho \leq 0.87$, all $p < 0.001$, Table 2). EPS changes over 3 years correlated with EDSS changes over 3 years ($0.39 \leq \rho \leq 0.47$, all $p \leq 0.004$, Table 3).

Predictive correlations between baseline EPS, short term changes in EPS and clinical outcome

Each baseline EPS was highly correlated to the $EDSS_{T4}$ ($0.84 \leq \rho \leq 0.86$, all $p < 0.001$) and moderately to $\Delta EDSS_{T4-T0}$ ($0.44 \leq \rho \leq 0.46$, $p < 0.001$, Table 4). Short-term changes over the first 6 months as detected by q-EPS correlated with both measures of clinical outcome, $EDSS_{T4}$ ($\rho = 0.56$, $p < 0.001$) and $\Delta EDSS_{T4-T0}$ ($\rho = 0.56$, $p < 0.001$). In contrast, no significant correlations were found for the short-term changes of the o-EPS and sq-EPS scores with clinical outcome (Table 4).

Predictive regression models on clinical outcome

The detailed results of the multivariable regression analysis on $EDSS_{T4}$ for each EPS are summarized in Table 5. For all three EPS, the backward selection resulted in models containing only the respective baseline EPS variable and $EDSS_{T0}$ with an adjusted R^2 between 0.85-0.86. The BICs were smaller (more negative) for the models based on either q-EPS or sq-EPS than for the model based on o-EPS, indicating a better model fit. The adjusted R^2 for a model based on $EDSS_{T0}$ alone was 0.77.

Figure 1 A-C shows the ROC curves for the prediction models of clinical worsening over 3 years (as defined in the methods section) based on the predictors EPS_{T0} , ΔEPS_{T1-T0} and their combination. The area under the curve (AUC) for each ROC curve is given in the legend. The results of the pairwise comparisons between the AUCs are summarized in Table 6. The AUC of the prediction model based on $\Delta q-EPS_{T1-T0}$ was significantly higher than the AUCs of the models based on the short-term change assessed by the other two scoring systems. However, no significant differences according to type of EPS were found for the AUCs of the models based

on the baseline variables alone or the combination of baseline variables and short-term change variables.

Discussion

All three EP scoring systems, no matter whether based on *ordinal*, *semi-quantitative* or *quantitative* scales, provided scores that consistently demonstrated strong cross-sectional and moderate longitudinal associations with clinical disability in MS.

Moreover, baseline EP scores, irrespectively of their scaling, consistently showed moderate to strong associations with both measures of medium-term clinical outcome, the EDSS at year 3 and the change in EDSS over 3 years. The combination of EPS and EDSS at baseline explained together 85-86% of EDSS variance at year 3.

Our results indicate that multimodal EP, irrespectively of the chosen scaling level robustly capture clinically relevant information in both cross-sectional and longitudinal analyses. Moreover, baseline EP are able to measure subclinical disease burden in MS that is predictive of later clinical outcome as shown by the independent contribution of baseline EPS to the prediction of EDSS at year three in a model that also contained baseline EDSS. This result indicates the potential of multimodal EP to help identify patients at risk for clinical worsening beyond what is possible on clinical information alone. However, short-term EP changes over 6 months did not improve overall prediction of EDSS at year 3 in multivariable regression analysis indicating that most of the information is already contained by the respective baseline EPS and baseline EDSS.

Our findings are in line with earlier studies that investigated combined EP in independent MS cohorts and predicted clinical outcome over 2 – 20 years using either *ordinal* scores (Leocani et al., 2006; Kallmann et al., 2006; Invernizzi et al.,

2011; Ramanathan et al., 2013), *semi-quantitative* scores (Jung et al., 2008; Margaritella et al., 2012a and b) or *quantitative* scores (Fuhr et al., 2001, Schlaeger et al., 2014b).

The *ordinal* scoring procedure has the advantage of being applicable without calculations and of being close to clinical routine. However, judgments of morphological abnormalities of potentials rely on the rater's expertise and are subjective, which may reduce reliability in particular in the context of multi-centric investigations. Moreover, individual alterations within the normal limits of a larger population are lost for the analysis.

The *semi-quantitative* scoring system is based on a stepwise classification of numerical abnormalities of evoked potential latencies and amplitudes. In contrast to the quantitative scoring method it includes an assessment of the presence or absence of side differences of amplitudes. However, the resulting score doesn't get beyond ordinal numbers, which implies a loss of information (Emerson, 1998).

The *quantitative* scoring system is based on the quantitative assessment of latencies only. Amplitudes are not included in the scoring procedure. Measurements of latencies also contain certain elements of subjectivity; however have been shown to be more reliably measurable in comparison to amplitudes (Comi et al, 1999).

In this study quantitatively and semi-quantitatively assessed EP changes were already significant after 6 months, when no significant changes were yet observed on the ordinal scale. The sample size needed to detect a significant short-term EPS change was smallest for the *quantitative* scaling system.

Moreover quantitatively assessed short-term EP changes demonstrated a more accurate prediction of clinical worsening in bivariate logistic regression analysis compared to short-term changes measured by other scaling systems and a significant positive correlation with later clinical outcome that was not found for the

other scaling systems. This is indicative of a higher sensitivity for clinically relevant short-term change (responsiveness) of the *quantitative* scoring system. The higher responsiveness is possibly due to a higher dynamic range as the approach based on z-values allows detection of changes within normal limits and within a broad range of pathological prolonged latencies. Models based on either quantitative or semi-quantitative scales, also consistently demonstrated a better model fit as indicated by a lower BIC in equally sized models than models based on ordinal scales.

These features are important, as the evaluation of short-term changes is crucial for clinical phase 2 trials. Ideally a surrogate (Strimbu et Travel, 2010) should exhibit a high sensitivity to change that becomes clinically meaningful later on.

In the subgroup of patients with primary progressive MS, the quantitative EPS was already shown to have a higher sensitivity for short-term change as compared to clinical measures of disability (EDSS) (Schlaeger et al., 2014).

Several limitations of this study have to be kept in mind when interpreting the aforementioned results.

A latency prolongation is not entirely specific for demyelination, nor is an amplitude reduction for axonal loss. For instance, a loss of axons can lead to an increase of latencies due to temporal dispersion requiring prolonged temporal summation in order to reach the depolarization threshold. In line with this pathophysiological consideration, several studies showed significant associations between reductions of retinal nerve fiber layer thickness (a measure reflecting axonal loss) and mfVEP latency delay in non-optic neuritis eyes of MS patients (Klistorner et al., 2013 and Sriram et al., 2014) indicating that the measurement of latencies may also, at least to some extent, reflect axonal loss.

On the other hand, amplitudes may be reduced due to temporal dispersion and phase cancellation and may recover in the presence of ongoing axonal degeneration (Klistorner et al., 2010) in particular in patients with evidence of remyelination. Consequently, amplitudes and latencies do not represent independent information. However, latency measurements are more reliable than amplitude measurements (Comi et al., 1999), which might explain their higher sensitivity to clinical relevant change in longitudinal assessments. Our data suggests that an increase in latency (that is measured during the relapse free interval) is clinically relevant, while longitudinal alterations in amplitudes did not show associations with the clinical evolution in our cohort.

A further limitation of the study is its retrospective design. The quantitative score was already tested on the same data set before (Schlaeger et al. 2012a, 2014a), and therefore, q-EPS might have had an accidental advantage over the other scoring systems. To provide full generalizability of the quantitative method, in particular in PPMS, our results have to be validated prospectively in an independent sample. However, the quantitative approach has been developed earlier in an independent cohort of RRMS and secondary progressive MS patients (Fuhr et al., 2001).

As done in precursor studies and in analogy to earlier z-transformation based studies on outcome measures in MS (Fischer et al., 1999, Fuhr et al., 2001) we substituted missing values in case of non-identifiable cortical components by the longest central conduction time observed in the whole cohort and within the same EP modality.

Although scores for the patients without reproducible response, therefore, are no longer quantitative *sensu stricto*, this scoring procedure seems to be the best compromise in order to not lose sensitivity by excluding data of patients with the most pathological results. We additionally also tested other strategies including (1) last observation carried forward, and (2) exclusion of curves with non-measurable

data. Regardless of the methodology chosen, the quantitative approach was most sensitive in detecting differences in evoked potentials over time.

A further limitation is inherent to the progressive course of the disease.

With regard to the electrophysiological examinations there was a selective dropout at later time points of patients with increasing disability mostly due to loss of mobility and the need for an accompanying person. The apparent ceiling effect of EPS at later time points, therefore, can be well explained by data loss of patients with high and fast changing EPS. Such dropouts pose difficulties in the analysis of clinical trials but are not specific to EP and relate to cohort studies in general. However, we were able to obtain the EDSS / EDSS by phone at year 3 in 74% of these dropouts (in 93% of all participants) so that their data was not lost for the predictive analyses.

The EPS in the current study are based on combined assessments of VEP, MEP to upper and lower extremities and median and tibial nerve SEP. Tibial SEP and MEP to the upper extremities have been shown previously to be most informative in the sub-sample of PPMS patients (Schlaeger et al. 2014a). Prediction of disability might be further improved by adding combined brainstem reflex responses and brainstem evoked potentials (Magnano et al., 2014) as well as vestibular myogenic evoked potentials (Ivancovic et al., 2013) to better capture brainstem involvement. Efficiency of the tests as well as the patient's ability and willingness to perform several additional examinations has to be taken into account when selecting EP modalities to optimize the evaluation of the disease course. However, the purpose of the present study was not to define the optimal selection of electrophysiological tests to characterize the course of MS, but to compare different methods to evaluate EP.

In conclusion, the present results support the validity of combined EP as monitoring tool independently from the chosen method of assessment, and therefore, the robustness of EP study results. The baseline variables derived by all three scoring tools show similar strong cross-sectional and moderate longitudinal correlations with disability. EPS at baseline independently improve prediction of EDSS after three years compared to baseline EDSS alone and therefore may help to better identify patients at risk for clinical worsening. The quantitative EPS, however, shows higher sensitivity to clinically relevant short-term changes than ordinal or semi-quantitative measures, most probably, as individual changes within the "normal" range can also be quantified.

Regardless of the lack of pathophysiological specificity, the scoring systems investigated in this study have produced a single score that correlates well with disease progress, allows disability to be monitored and might serve as a selection criterion and surrogate outcome for the disease course in MS trials.

Conflict of interest

None of the authors have potential conflicts of interest with this study to be disclosed.

Acknowledgements

We thank Sylvia Fluri, Monika Kungler, Susanne Tanner, Beatrice Wessner and Jacqueline Fleury for technical assistance. The study has been supported by the Swiss National Science Foundation, grant SPUM 33CM30-140338, and by the Swiss Multiple Sclerosis Society.

References

Calabresi PA, Radue EW, Goodin D, Jeffery D, Rammohan KW, Reder AT, et al. Safety and efficacy of fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS II): a double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet Neurol* 2014;13:545-556.

Celesia GG, Brigell MG. Recommended standards for pattern electroretinograms and visual evoked potentials. *Electroencephalogr Clin Neurophysiol Suppl* 1999;52:53-67.

Comi G, Leocani L, Medaglini S, Locatelli T, Martinelli V, Santuccio G, Rossi P. Measuring evoked responses in multiple sclerosis. *Mult Scler* 1999;5:263-267.

Emerson RG. Evoked potentials in clinical trials for multiple sclerosis. *J Clin Neurophysiol* 1998;15:109-116.

Fischer JS, Rudick RA, Cutter GR, Reingold SC. The Multiple Sclerosis Functional Composite Measure (MSFC): an integrated approach to MS clinical outcome assessment. National MS Society Clinical Outcomes Assessment Task Force. *Mult Scler* 1999;5:244-250.

Fuhr P, Borggreffe-Chappuis A, Schindler C, Kappos L. Visual and motor evoked potentials in the course of multiple sclerosis. *Brain* 2001; 124:2162-2168.

Invernizzi P, Bertolasi L, Bianchi MR, Turatti M, Gajofatto A, Benedetti MD.

Prognostic value of multimodal evoked potentials in multiple sclerosis: the EP score.

J Neurol 2011; 258:1933-1939.

Ivanković A, Neseck Mađarić V, Starčević K, Krbot Skorić M, Gabelić T, Adamec I, Habek M. Auditory evoked potentials and vestibular evoked myogenic potentials in evaluation of brainstem lesions in multiple sclerosis. J Neurol Sci. 2013;328:24-27.

Jung P, Beyerle A, Ziemann U. Multimodal evoked potentials measure and predict disability progression in early relapsing-remitting multiple sclerosis. Mult Scler 2008; 14:553-556.

Kallmann BA, Fackelmann S, Toyka KV, Rieckmann P, Reiners K. Early abnormalities of evoked potentials and future disability in patients with multiple sclerosis. Mult Scler 2006; 12:58-65.

Klistorner A, Arvind H, Garrick R, Graham SL, Paine M, Yiannikas C.

Interrelationship of optical coherence tomography and multifocal visual-evoked potentials after optic neuritis. Invest Ophthalmol Vis Sci 2010; 51:2770-2777.

Klistorner A, Garrick R, Barnett MH, Graham SL, Arvind H, Sriram P, et al. Axonal loss in non-optic neuritis eyes of patients with multiple sclerosis linked to delayed visual evoked potential. Neurology 2013; 80:242-245.

Kappos L, Lechner-Scott J, Wu S. Neurostatus.net. 2007. Available at:

<http://www.neurostatus.net>

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

Lechner-Scott J, Kappos L, Hofman M, Polman CH, Ronner H, Montalban X, et al. Can the Expanded Disability Status Scale be assessed by telephone? *Mult Scler* 2003; 9:154-159.

Leocani L, Rovaris M, Boneschi FM, Medaglini S, Rossi P, Martinelli V, Amadio S, et al. Multimodal evoked potentials to assess the evolution of multiple sclerosis: a longitudinal study. *J Neurol Neurosurg Psychiatry* 2006; 77:1030-1035.

Magnano I, Pes GM, Pilurzi G, Cabboi MP, Ginatempo F, Giaconi E, Tolu E, Achene A, Salis A, Rothwell JC, Conti M, Deriu F. Exploring brainstem function in multiple sclerosis by combining brainstem reflexes, evoked potentials, clinical and MRI investigations. *Clin Neurophysiol*. 2014;125:2286-2296.

Mauguière F, Allison T, Babiloni C, Buchner H, Eisen AA, Goodin DS, et al. Somatosensory evoked potentials. The International Federation of Clinical Neurophysiology. *Electroencephalogr Clin Neurophysiol Suppl* 1999;52:79-90.

Margaritella N, Mendozzi L, Garegnani M, Nemni R, Colicino E, Gilardi E, et al. Exploring the predictive value of the evoked potentials score in MS within an appropriate patient population: a hint for an early identification of benign MS? *BMC Neurology* 2012a; 12:80.

Margaritella N, Mendozzi L, Garegnani M, Colicino E, Gilardi E, Deleonardis L, et al. Sensory evoked potentials to predict short-term progression of disability in multiple sclerosis. *Neurol Sci* 2012b; 33:887-892.

McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD, et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the international panel on the diagnosis of multiple sclerosis. *Ann Neurol* 2001; 50:121–127.

O'Connor P, Marchetti P, Lee L, Perera M. Evoked potential abnormality scores are a useful measure of disease burden in relapsing-remitting multiple sclerosis. *Ann Neurol* 1998; 44:404-407.

Ramanathan S, Lenton K, Burke T, Gomes L, Storchenegger K, Yiannikas C, et al. The utility of multimodal evoked potentials in multiple sclerosis prognostication. *J Clin Neurosci* 2013; 20:1576-1581.

Rothwell JC, Hallett M, Berardelli A, Eisen A, Rossini P, Paulus W. Magnetic stimulation: motor evoked potentials. The International Federation of Clinical Neurophysiology. *Electroencephalogr Clin Neurophysiol Suppl* 1999; 52:97-103.

Schlaeger R, D'Souza M, Schindler C, Grize L, Kappos L, Fuhr P. Combined evoked potentials as markers and predictors of disability in early multiple sclerosis. *Clin Neurophysiol* 2012a; 123:406-410.

Schlaeger R, D'Souza M, Schindler C, Grize L, Kappos L, Fuhr P. Prediction of long-term disability in multiple sclerosis. *Mult Scler* 2012b; 18:31-38.

Schlaeger R, D'Souza M, Schindler C, Grize L, Kappos L, Fuhr P.
Electrophysiological markers and predictors of the disease course in primary progressive multiple sclerosis. *Mult Scler* 2014a; 20:51-56.

Schlaeger R, Schindler C, Grize L, Kappos L, Fuhr P. Combined evoked potentials predict MS disability after 20 years. *Mult Scler* 2014b; 20:1348-1354.

Sriram P, Wang C, Yiannikas C, Garrick R, Barnett M, Parratt J, et al. Relationship between optical coherence tomography and electrophysiology of the visual pathway in non-optic neuritis eyes of multiple sclerosis patients. *PLoS One*. 2014;9:e102546.

Stoehr M. Somatosensible Reizantworten von Nerven, Rückenmark und Gehirn. In Stoehr M, Dichgans J, Buettner UW, Hess CW ed. *Evozierte Potentiale*. 3rd edition. Heidelberg: Springer Verlag; 1996; p. 72;625-627.

Strimbu K, Tavel JA. What are Biomarkers? *Curr Opin HIV AIDS*. 2010;5:463-466.

Thompson AJ, Montalban X, Barkhof F, Brochet B, Filippi M, Miller DH, et al.
Diagnostic criteria for primary progressive multiple sclerosis: a position paper. *Ann Neurol* 2000; 47:831-835.

Figure Legends

Figure 1. Receiver operating characteristic (ROC) curves showing the sensitivity and specificity for the prediction of clinical worsening over 3 years given by binary logistic models based on either (A) the baseline variables q-EPS_{T0}, sq-EPS_{T0}, o-EPS_{T0} as single predictors, (B) the corresponding variables of short-term change within the first 6 months Δ q-EPS_{T1-T0}, Δ sq-EPS_{T1-T0}, and Δ o-EPS_{T1-T0}, respectively, and (C) the combinations of the baseline and corresponding variable of short-term change, respectively. Worsening thereby was defined as an increase in EDSS by 1.5 points or more if the baseline EDSS was 0, by 1.0 point or more if the baseline EDSS was between 1 and 5.0 and by 0.5 points otherwise.

Table 1

Distribution of EP-Scores at each time point

	Baseline (T0)		Month 6 (T1)		Month 12 (T2)		Month 24 (T3)		Month 36 (T4)	
	Mean/ SD	Min/Max	Mean/ SD	Min/Max	Mean/ SD	Min/Max	Mean/ SD	Min/Max	Mean/ SD	Min/Max
N	72 (50)		66 (50)		60 (50)		55 (50)		53 (50)	
q-EPS	2.99/3.62	-0.75/18.07	3.53/4.41	-0.95/20.11	3.61/4.06	-0.83/15.44	3.57/4.11	-0.72/15.35	3.58/3.91	-0.59/12.71
	<i>2.59/2.69</i>	<i>-0.75/8.66</i>	<i>3.20/3.70</i>	<i>-0.71/16.45</i>	<i>3.49/3.94</i>	<i>-0.83/15.43</i>	<i>3.61/3.94</i>	<i>-0.72/15.35</i>	<i>3.51/3.90</i>	<i>-0.59/12.71</i>
o-EPS	0.65/0.57	0/2.3	0.68/0.62	0/2.2	0.67/0.58	0/1.7	0.69/0.57	0/1.9	0.71/0.56	0/1.7
	<i>0.60/0.49</i>	<i>0/1.7</i>	<i>0.64/0.58</i>	<i>0/2.1</i>	<i>0.64/0.54</i>	<i>0/1.7</i>	<i>0.72/0.56</i>	<i>0/1.9</i>	<i>0.69/0.55</i>	<i>0/1.7</i>
sq-EPS	1.25/1.08	0/4.2	1.32/1.15	0/4.2	1.36/1.13	0/3.9	1.30/1.11	0/3.5	1.38/1.12	0/3.6
	<i>1.14/0.88</i>	<i>0/2.9</i>	<i>1.22/1.00</i>	<i>0/3.6</i>	<i>1.28/1.02</i>	<i>0/3.9</i>	<i>1.34/1.09</i>	<i>0/3.5</i>	<i>1.35/1.09</i>	<i>0/3.6</i>

EP: Evoked Potentials; T0: baseline, T1: month 6, T2: year1, T3: year2, T4: year3; q-EPS: quantitative EP-Score, o-EPS: ordinal EP-Score, sq-EPS: semi-quantitative EP-Score.

Electrophysiological data at the time-points T0 to T4 of all participants that were investigated at the respective time-point (first line) and of those with complete electrophysiological and clinical data at all time-points (in italics). With regard to the electrophysiological examination there was a selective dropout of patients with severe disease courses at later time-points. Patients, who dropped out of the electrophysiological part of the study, were investigated either clinically or by phone, if possible (see methods).

Table 2

Cross-sectional correlations between EP-Scores and EDSS

Score	EDSS_{T0}			EDSS_{T1}			EDSS_{T2}			EDSS_{T3}			EDSS_{T4}		
	rho	p	95% CI	rho	p	95% CI	rho	p	95% CI	rho	p	95% CI	rho	p	95% CI
N	72			66			60			55			53		
q-EPS	0.81	<0.001	0.72-0.87	0.83	<0.001	0.74-0.89	0.85	<0.001	0.77-0.91	0.79	<0.001	0.65-0.87	0.87	<0.001	0.78-0.92
o-EPS	0.81	<0.001	0.71-0.87	0.83	<0.001	0.75-0.88	0.80	<0.001	0.71-0.87	0.73	<0.001	0.59-0.83	0.81	<0.001	0.71-0.88
sq-EPS	0.83	<0.001	0.74-0.88	0.81	<0.001	0.71-0.87	0.79	<0.001	0.69-0.86	0.72	<0.001	0.58-0.81	0.79	<0.001	0.68-0.86

EP: Evoked Potentials. EDSS: Expanded Disability Status Scale. Rho: Spearman rank rho; T0: baseline, T1: month 6, T2: year1, T3: year2, T4: year3; q-EPS: quantitative EP-Score, o-EPS: ordinal EP-Score, sq-EPS: semi-quantitative EP-Score. 95%-Confidence intervals (CI) were calculated using bootstrapping with 1000 iterations.

Table 3

Longitudinal correlations between EP-Scores and EDSS over 3 years

Score	Δ EDSS _{T4-T0}		
	rho	p	95% CI
Δ q-EPS _{T4-T0}	0.47	<0.001	0.25-0.64
Δ o-EPS _{T4-T0}	0.41	0.003	0.14-0.62
Δ sq-EPS _{T4-T0}	0.39	0.004	0.16-0.58

EP: Evoked Potentials. EDSS: Expanded Disability Status Scale.

Spearman rank rho; T0: baseline, T4: year 3; q-EPS: quantitative EP-Score,

o-EPS: ordinal EP-Score, sq-EPS: semi-quantitative EP-Score.

95%-Confidence intervals (CI) were calculated using bootstrapping with 1000 iterations.

Rho:

Table 4

Predictive correlations between EP scores at baseline and clinical development over 3 years

Score	EDSS_{T4}			Δ EDSS_{T4-T0}		
	rho	p	95% CI	rho	p	95% CI
q-EPS_{T0}	0.85	≤ 0.001	0.76-0.91	0.46	≤ 0.001	0.25-0.64
o-EPS_{T0}	0.84	≤ 0.001	0.77-0.89	0.44	≤ 0.001	0.22-0.62
sq-EPS_{T0}	0.86	≤ 0.001	0.79-0.90	0.44	≤ 0.001	0.22-0.63
Δq-EPS_{T1-T0}	0.56	≤ 0.001	0.34-0.73	0.56	≤ 0.001	0.37-0.71
Δo-EPS_{T1-T0}	0.21	0.091	-0.09-0.47	0.16	0.197	-0.10-0.42
Δsq-EPS_{T1-T0}	0.15	0.234	-0.07-0.36	0.30	0.014	0.08-0.51

EP: Evoked Potentials. EDSS: Expanded Disability Status Scale. Rho: Spearman rank rho; T0: baseline, T1: month 6, T4: year 3; q-EPS: quantitative EP-Score, o-EPS: ordinal EP-Score, sq-EPS: semi-quantitative EP-Score. 95%-Confidence intervals (CI) were calculated using bootstrapping with 1000 iterations.

Table 5

Multivariable regression analysis examining the predictive relationship of baseline EPS variables, their short-term change and other baseline characteristics with clinical outcome at year 3. Models 1a-c were reduced in a backward selection process to Models 2a-c.

Model	Outcome	Predictors	Parameter Estimates	Std.Err.	p
1a R ² =0.87 adj R ² =0.86 BIC=-42.03	EDSS _{T4}	Intercept q-EPS _{T0} Δ q-EPS _{T1-T0} EDSS _{T0} Age Sex	0.367 0.157 0.080 0.760 0.009 -0.042	0.366 0.038 0.075 0.096 0.010 0.192	0.319 <0.001 0.289 <0.001 0.365 0.826
2a R ² =0.86 adj R ² =0.86 BIC=-46.98	EDSS _{T4}	Intercept q-EPS _{T0} EDSS _{T0}	0.528 0.174 0.814	0.183 0.033 0.081	0.002 <0.001 <0.001
1b R ² =0.86 adj R ² =0.85 BIC=-37.28	EDSS _{T4}	Intercept o-EPS _{T0} Δ o-EPS _{T1-T0} EDSS _{T0} Age Sex	0.328 1.093 0.227 0.737 0.007 -0.082	0.378 0.268 0.493 0.108 0.010 0.199	0.389 <0.001 0.646 <0.001 0.485 0.683
2b R ² =0.86 adj R ² =0.85 BIC=-43.25	EDSS _{T4}	Intercept o-EPS _{T0} EDSS _{T0}	0.473 1.142 0.771	0.184 0.242 0.093	0.012 <0.001 <0.001
1c R ² =0.87 adj R ² =0.86 BIC=-42.89	EDSS _{T4}	Intercept sq-EPS _{T0} Δ sq-EPS _{T1-T0} EDSS _{T0} Age Sex	0.098 0.621 0.335 0.701 0.012 -0.006	0.376 0.143 0.267 0.109 0.010 0.194	0.795 <0.001 0.216 <0.001 0.221 0.974
2c R ² =0.86 adj R ² =0.86 BIC=-46.65	EDSS _{T4}	Intercept sq-EPS _{T0} EDSS _{T0}	0.490 0.664 0.733	0.179 0.128 0.093	0.008 <0.001 <0.001

EP: Evoked Potentials. T0: baseline, T1: month 6, T4: year 3; q-EPS: quantitative EP-Score, o-EPS: ordinal EP-Score, sq-EPS: semi-quantitative EP-Score. BIC: Bayesian information criterion. Parameter estimate relates to the respective parameter unit. Adj.: adjusted. EDSS: Expanded Disability Status Scale. Std.Err.: Standard Error.

Table 6

Pairwise comparison between the areas under the receiver operator characteristic curve for the prediction of clinical worsening over 3 years based on the different EP scoring systems using baseline variables, variables of short-term change and their respective combinations.

Comparison	AUC_1	AUC_2	Difference= AUC_1 - AUC_2	
	Mean $\pm$ StdDev	Mean $\pm$ StdDev	Mean $\pm$ StdDev	p-value
Baseline only q-EPS versus o-EPS	0.7439 $\pm$ 0.0646	0.7484 $\pm$ 0.0632	-0.0045 $\pm$ 0.0261	0.878
Baseline only q-EPS versus sq-EPS	0.7439 $\pm$ 0.0646	0.7410 $\pm$ 0.0637	0.0029 $\pm$ 0.0204	0.880
Baseline and change q-EPS versus o-EPS	0.7871 $\pm$ 0.0662	0.7462 $\pm$ 0.0662	0.0410 $\pm$ 0.0537	0.456
Baseline and change q-EPS versus sq-EPS	0.7871 $\pm$ 0.0642	0.7619 $\pm$ 0.0609	0.052 $\pm$ 0.0415	0.544
Change only q-EPS versus o-EPS	0.7865 $\pm$ 0.0589	0.5592 $\pm$ 0.0575	0.2293 $\pm$ 0.0817	0.004
Change only q-EPS versus sq-EPS	0.7865 $\pm$ 0.0589	0.6173 $\pm$ 0.0681	0.1692 $\pm$ 0.0692	0.014

AUC: Area under the ROC curve. StdDev: Standard deviation. T0: baseline, T1: month 6. q-EPS: quantitative EP-Score, o-EPS: ordinal EP-Score, sq-EPS: semi-quantitative EP-Score.

Baseline only: Comparison between ROC curves based on models with EPS_{T0} as single predictor.

Baseline and change: Comparison between ROC curves based on models with EPS_{T0} and short-term change Δ EPS_{T1-T0} as combined predictors.

Change only: Comparison between ROC curves based on models with short-term change Δ EPS_{T1-T0} as single predictor.

Calculations were done using bootstrapping with 1000 times re-sampling.