Original Research Paper

Neurostatus e-Scoring improves consistency of Expanded Disability Status Scale assessments: A proof of concept study

Marcus D’Souza, Özgür Yaldızli, Roland John, Deborah R Vogt, Athina Papadopoulou, Elisabeth Lucassen, Milena Menegola, Michaela Andelova, Frank Dahlke, Franz Schnyder and Ludwig Kappos

Abstract

Background: To improve the consistency of standardized Expanded Disability Status Scale (EDSS) assessments, an electronic data capture tool and analysis tool was developed, Neurostatus e-Scoring (NESC). This tool allows real-time feedback by comparing entries with established scoring rules.

Objective: To test whether using NESC reduces inconsistencies as compared to the paper-and-pencil version of the Expanded Disability Status Scale (pEDSS).

Methods: In all, 100 multiple sclerosis (MS) patients were assessed in random order on the same day by pairs of neurologists, one using pEDSS and one NESC. We compared inter-rater reliability and frequency of inconsistencies in Neurostatus subscores, functional system (FS) scores, ambulation and EDSS steps.

Results: Inconsistencies of any type were more likely to occur when using pEDSS (mean odds ratio (95% confidence interval (CI)) = 2.93 (1.62; 5.29)). This was also the case for FS score inconsistencies (2.54 (1.40; 4.61)) and more likely for patients in the lower EDSS range (≤3.5 vs >3.5) (5.32 (1.19; 23.77)). Overall, inter-rater agreement for the assessed Neurostatus subscores was high (median and inter-quartile range = 0.84 (0.73, 0.81)).

Conclusion: Our data provide class II evidence that the use of NESC increases consistency of standardized EDSS assessments, and may thus have the potential to decrease noise and increase power of MS clinical trials.

Keywords: Multiple sclerosis, EDSS, neurostatus, clinical trial, disability

Date received: 9 March 2016; revised: 4 May 2016; accepted: 8 June 2016

Introduction

The clinical assessment of impairment and disability in multiple sclerosis (MS) remains the most important outcome measure in daily care and in therapeutic trials. Although frequently criticized, the Expanded Disability Status Scale (EDSS) and its functional systems (FS) are the most widely used clinical outcome measure and the only scale accepted by most health agencies for approval of MS therapeutics.1,2 Since its introduction by John Kurtzke3 first in 1955 as the Disability Status Scale (DSS), then modified in 19834 as the EDSS, its low inter- and intra-rater reliability has been a source of concern.5-11 This is partially due to ambiguous rules for the assignment of the seven FS scores and the EDSS.

In an attempt to improve reliability and internal consistency of assessments, a set of scoring rules that were based on the findings of a standardized neurological examination, the ‘Neurostatus’ was developed in the early 1990s.12 Based on expert consensus, this system retains the main principles of the original scale but deals with some of its ambiguities by providing more explicit rules and definitions. Neurostatus consists of three steps: (a) a standardized semi-quantitative neurological examination of all signs and symptoms including standardized questions about functions that cannot be assessed fully by the neurological examination and the assessment of ambulation, (b) assessment of the FS scores based on this documented comprehensive examination and a set of definitions and scoring criteria and (c) calculation of the EDSS step, based on the FS scores and the assessment of walking range, need for walking assistance or support in daily activities (graded with the Ambulation score (AS)).13,14 To achieve a higher reliability and to
increase consistency of the assessments, definitions and scoring rules are provided in written format and users can benefit from interactive audio-visual investigator training sessions and a teaching DVD that contains video clips with examples for the different disability grades. To document mastery of the ‘Neurostatus-EDSS’, a web-based certification tool (Neurostatus e-Test) was developed.\textsuperscript{15}

As of today, the ‘Neurostatus-EDSS’ has been adopted as standard in more than 120 phase II and III clinical trials in MS, including most of the pivotal trials that have led to the approval of current MS treatments. Since its introduction, more than 10,000 investigators have been certified through the Neurostatus e-Test certification process. Introducing detailed definitions and rules helped to define common standards of scoring, but in a disease with such a wide spread symptomatology and severity, these detailed rules and definitions have also increased the complexity of assessments and thus the probability of inconsistent ratings and scores.\textsuperscript{1}

Neurostatus e-Scoring (NESC) was developed to further improve the reliability and internal consistency of EDSS assessments. It is an electronic data capture, analysis and management system, which allows capturing of the results of the neurological examination including the Neurostatus subscores, FS scores, the AS and the final EDSS step, with a touch screen based tablet device (iPad) and provides algorithm-based real-time feedback on potential inconsistencies. Neurostatus Systems AG validated the NESC system according to the regulatory requirements and maintains the validated state of the system.

In this prospective randomized study, we investigated the hypothesis that fewer inconsistencies occur in the calculation of the FS scores and EDSS steps when using NESC as compared to the standardized paper-and-pencil EDSS version (pEDSS).

Materials and methods

NESC algorithm

The NESC algorithm is based on the definitions provided in the ‘Neurostatus’ for the standardized assessment of Neurostatus subscores, FS scores, the AS and the final EDSS step.\textsuperscript{13,14} It compares the Neurostatus subscores against the FS scores as well as the FS scores together with the AS against the final EDSS step. Although an automatic calculation for each of the 7 FS scores and the final EDSS step is implemented, the algorithm does not automatically provide the correct scores but instead provides feedback about possible inconsistencies and calculation errors from the entered source data and the calculations of the physician. The final decision on the definitive scores remains with the physician, who can overrule the algorithm at any stage but has to explain his decision using a comment function.

Prior to programming the NESC algorithm, all possible combinations of each FS as derived from the corresponding Neurostatus subscores were summarized in a table. Lower and upper limits of each FS score of the 7 FS, the AS and the final EDSS step were defined based on the Neurostatus definitions developed for the use in clinical trials.\textsuperscript{13,14} Furthermore, cross checks using overlapping or redundant items of the Neurostatus, FS and EDSS scoring system (e.g. muscle strength or ataxia of the lower extremities against ambulation tested) were integrated. After validation on a test sample of 115 examinations (not part of this clinical study), the algorithm was implemented on an iPad App. The user interface of the iPad allows data capture with a touch screen (Figure 1) that was designed based on the paper-and-pencil scoring sheets used in clinical trials.

Study design and population

The local ethics committee (Ethikkommission beider Basel (EKBB)) had no concerns regarding this study. Patients attending the MS outpatient clinic for regularly scheduled visits were asked for consent to undergo the neurological examination twice on the same day. All patients were diagnosed with MS according to the McDonald criteria and had an EDSS between 0 and 9. Three neurologists with experience in MS and trained in the Neurostatus assessment were instructed in the use of an iPad device and the technical functions of NESC. After written informed consent, the MS patients were assessed by two of the three neurologists independently at the same day, one using the pEDSS and the other NESC. After a full neurological examination, the Neurostatus subscores, FS scores, AS and the final EDSS step were assessed based on the Neurostatus definitions\textsuperscript{13,14} and either documented on the iPad (touch screen) or on a paper scoring sheet. To reduce the fluctuations of disease symptoms from day to day, both examinations had to be performed on the same day. Furthermore, to control for bias related to potential fluctuations in patient performance within a day or due to fatigue/training effect caused by the preceding examination, the assessments were performed right after each other and the order of the methods was randomized. The assessing neurologists were blinded to the findings.
and scores of each other’s assessments. In a further step, an independent and EDSS experienced neurologist checked manually for inconsistencies in the entered/written Neurostatus subscores, FS scores, the AS and the final EDSS step at study end. This consistency check was performed twice on printouts, blinded for the method of assessment (pEDSS vs NESC) and to the identities of the examining neurologists and patients, strictly based on the current Neurostatus definitions and scoring table.\(^{13,14}\) Furthermore, the Neurostatus subscore entries were compared for pairs of assessments of the same patients to determine the inter-rater agreement on different parts of the neurological examination (Neurostatus subscores) irrespectively of the method used (pEDSS or NESC).

**Statistics**

NESC and pEDSS were compared for the number of inconsistencies in the calculation of FS scores, AS and EDSS steps as determined by the independent neurologist. FS score inconsistencies were defined as FS scores that were not assigned in accordance to the respective Neurostatus subscores. EDSS step inconsistencies were defined as EDSS steps that were not assigned in accordance to the combination of respective FS scores and AS (irrespective of whether derived calculations of the FS scores or the AS themselves were correct). Furthermore, we evaluated the occurrence of those inconsistencies in FS scores, which if corrected would have resulted in a different EDSS step (EDSS step relevant FS inconsistencies). All endpoints were binary variables. Inconsistencies of any type, FS score inconsistencies and EDSS step relevant FS score inconsistencies were analysed using generalized mixed-effects models with a binomial error distribution. Patients were included as random effects to account for repeated measures. EDSS assessment method (NESC vs pEDSS) and examiner and level of EDSS (‘low’ = EDSS \(\leq 3.5\), ‘high’ = EDSS > 3.5) were considered as fixed effects. We used Akaike’s information criterion (AIC) to select for each endpoint the fixed effects structure with most empirical support. AS inconsistencies and EDSS step inconsistencies were too rare to apply logistic regression methods and were therefore analysed using the exact McNemar test. All effect sizes are expressed as odds ratios (OR) with 95% confidence intervals (CIs).

The assigned Neurostatus subscores from either NESC or pEDSS were compared for differences between the pairs of neurologists seeing the same patient on the same day. Since for many items, patients were rated ‘identically’ with very little observed disagreement – for example, only two out of four or five possible categories were used – standard measures of inter-rater reliability, such as Cohen’s \(\kappa\) or Krippendorff’s \(\alpha\), which are weighted for ‘expected disagreement’, are not reliable measures. Therefore, percentage agreement is presented.

**Results**

The patient’s characteristics are listed in Table 1. The frequencies of FS grades, AS and EDSS steps...
calculated by pEDSS and NESC are summarized in Supplementary Table 1.

For all endpoints, inconsistencies occurred more likely when using pEDSS compared to NESC (Figure 2). Examiners were three times as likely to provide inconsistent assessments of any type when using pEDSS compared to using NESC (mean OR (95% CI) = 2.93 (1.62; 5.29)). Also FS score inconsistencies occurred more often when pEDSS was used compared to NESC (2.54 (1.40; 4.61)), this difference being even more accentuated for EDSS step relevant FS score inconsistencies (4.13 (1.45; 11.76)). Furthermore, FS score inconsistencies were much more likely to occur for patients with lower EDSS steps (⩽3.5) than for patients with higher EDSS steps (>3.5) (5.32 (1.19; 23.77)). Inconsistencies regarding the AS (five errors in total: four with pEDSS and one with NESC) and the calculation of the EDSS step (eight errors in total: six with pEDSS and two with NESC) were rare and ORs were not significantly different between pEDSS and NESC for either AS inconsistencies (0.25 (0.01; 2.53)) or EDSS inconsistencies (0.33 (0.03; 1.86)). The occurrence of inconsistencies did not differ significantly between examiners, across different courses of MS (relapsing-remitting multiple sclerosis (RRMS), secondary progressive multiple sclerosis (SPMS) and primary progressive multiple sclerosis (PPMS)), or by sequence of methods applied (data not shown).

Median percentage agreement for all Neurostatus subscores was 0.84 (interquartile range (IQR): 0.73–0.81, range: 0.52–0.93). Lower agreement was found for the tendon reflexes (percentage agreement: 0.52–0.58), non-motor-fatigue (percentage agreement: 0.58) and vibration sense of the lower extremities (percentage agreement: 0.64–0.65), as well as for eye movements (percentage agreement: 0.64) and nystagmus (percentage agreement: 0.65). Higher agreement was found for muscle strength (percentage agreement: 0.81–0.97) and the position test for the upper extremities (percentage agreement: 0.99). Agreements for each Neurostatus subscore are reported in Figure 3 and in Supplementary Table 2.

### Table 1. Patient characteristics.

<table>
<thead>
<tr>
<th>Patient characteristics (n = 100)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (median, range), years</td>
<td>45 years (21–75)</td>
</tr>
<tr>
<td>Gender (female/male)</td>
<td>72 females/28 males</td>
</tr>
<tr>
<td>Disease duration (median, range), years</td>
<td>9 years (0–40)</td>
</tr>
<tr>
<td>Disease course (RRMS/SPMS/PPMS)</td>
<td>74 RRMS, 22 SPMS, 4 PPMS</td>
</tr>
<tr>
<td>EDSS (median, range)</td>
<td>3.0 (0–9)</td>
</tr>
</tbody>
</table>


### Figure 2. Total number of inconsistencies in scoring functional systems and the EDSS (FS) by assessment method. White columns represent NESC, grey columns pEDSS assessments.


**Discussion**

Variability of EDSS assessments can be due to daily and circadian fluctuations of MS symptoms, differences in the examination techniques and the experience of the neurologists interpreting MS symptoms, but also can be attributable to ambiguous grading rules. Daily fluctuations can hardly be avoided although probably best addressed using a standardized schedule and time of assessments. Differences in examination techniques and experience can be
addressed, at least to some degree, with standardized trainings (including re-trainings on a regular basis) and by avoiding using multiple EDSS examiners for a patient during a clinical trial. Variability due to ambiguous grading rules is also an important factor that can be addressed by training sessions and audio-visual training material.

The results of this study show a lower measurement variability and improved consistency of NESC over the pEDSS in calculating and assessing FS scores, AS and EDSS steps. This superiority was observed across all disease severity grades but was more pronounced in patients with low EDSS steps (≤3.5). In the lower part of the EDSS scale, the impact of the FS scores, which reflect the complete neurological examination, on the EDSS step is more pronounced and has been made responsible for the lower reliability at the lower range of the EDSS in comparison to higher range which is mainly defined by the ability to walk or the need of assistance for walking or daily activities.1,9,16 The higher error rate of the cerebral FS assessments (Figure 2) was mainly due to inconsistent instructions to the neurologists whether fatigue should be taken into consideration when assessing the cerebral FS score or not. Depending on the specific protocols, in some trials, the fatigue subscores are not taken into account for the cerebral FS score. Although the NESC was programmed to include fatigue when calculating the cerebral FS and to alert the investigator if this was not done the relative advantage of NESC is in line with the other FS scores.

The neurological examination is known to have a high variability depending on the experience of the examiner and the neurological sign being assessed.17,18 Additionally, the EDSS is known for its low inter-rater reliability on both the level of the EDSS steps and the FS scores.2,9,16,19 To our knowledge, the reliability of the assessment of complete sets of signs and symptoms (Neurostatus subscores) as constituents of the FS and EDSS scores has not been analysed to this date. In this study, we assessed inter-rater reliability only across methods due to feasibility considerations. Although our result is based on a low number of neurologists – all from one centre – the high inter-rater reliability between pairs of neurologists for all Neurostatus subscores is reassuring. This high inter-rater reliability across methods supports the value of the detailed Neurostatus definitions that were used for both NESC and pEDSS assessments. As a back side of such detailed definitions, in a disease like MS with
wide spread symptomatology and severity, increased complexity of assessments leads to increased probability of inconsistent ratings and scores. Our results clearly support the value of NESC in addressing this problem, thus contributing to more accurate and less noisy assessments. Further studies with more investigators and across different centres investigating intra- and inter-rater reliability of the assessment of different signs and symptoms and their impact on FS and EDSS scores would be desirable.

Although the EDSS examiners participating in this study were well trained and experienced in applying the Neurostatus definitions, they still clearly benefited from the NESC and its real-time feedback on inconsistencies in rating. Thus, one would assume that EDSS examiners with less experience with the Neurostatus-EDSS assessments would benefit even more from the NESC. As a result, the advantages of the NESC compared to the pEDSS in the setting of multicentre clinical trials and in clinical practice (including institutions that are not particularly specialized in MS) could be even higher than shown in our trial.

To date, the NESC algorithm has already been implemented in five global phase III multicentre studies. Three of these studies use the NESC algorithm together with a third party data capture and management system, and two use the full NESC system, including iPad for data capture, real-time feedback and a central review by an EDSS expert. First results with the latter suggest further improvements of the consistency of EDSS-data.

Several electronic EDSS calculators are available as download applications, for example, http://www.mikesiosapps.com, http://www.rodanotech.ch or http://www.torontoneuroapps.ca. Markowitz et al. performed a study using an electronic data capture device to compare paper and pencil assessments with an electronic version for calculation of the EDSS step based on the FS scores provided by the examiner only. Gaspari et al. developed an automatic ‘EDSS expert system’ which interacts with the examiner when EDSS >3.5 is assigned. Cohen et al. recommended a computer-aided instrument for neurological evaluation of patients with MS. They provided a guided interface for the neurological examination with a list of closed questions and included a structured ambulation questionnaire, as well as an automated FS score and EDSS step calculation. Further some calculators are provided for the direct use by patients, for example, http://multiple-sclerosis-research.blogspot.com, to track their own disability.

There is no information about test characteristics of these tools and their agreement with expert assessments. Most of the reported automatic calculation systems lack detailed information about the specific rules used to determine the FS scores and EDSS steps. Some provide the calculation of the EDSS step out of given FS scores and ambulation only. None of these tools provides guidance on how to determine the FS based on a complete clinical examination.

To conclude, in this study, the algorithm-based NESC was superior when compared with pEDSS in providing consistent assessments. Our results imply that the Neurostatus definitions improve inter-rater reliability and that the use of NESC for EDSS assessment improves consistency and thus further reduces measurement noise. Eventually, this should increase the quality and efficiency of clinical trials that use the Neurostatus-EDSS as an outcome.

Acknowledgements
Beat Gersbach and Viktor Del Ponte supported programming of the algorithm.

Declaration of Conflicting Interests
The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Marcus D’Souza received travel support from Bayer AG, Teva and Genzyme and research support from the University Hospital Basel; Özgür Yaldızlı has nothing to disclose; Deborah R. Vogt has nothing to disclose; Roland John is an employee of Neurostatus Systems AG; Athina Papadopoulos consulted for Teva, received travel grants from Bayer AG, Teva and UCB-Pharma AG and research support from the University of Basel; Elisabeth Lucassen has nothing to disclose; Milena Menegola has nothing to disclose; Michaela Andelova has nothing to disclose; Deborah Vogt has nothing to disclose; Michaela Andelova has nothing to disclose; Frank Dahlke is an employee of Novartis Pharma AG Basel; Franz Schnyder is an employee and owns stocks of Neurostatus Systems AG; Ludwig Kappos’ institution, the University Hospital Basel, has received research support and payments that were used exclusively for research support for Prof Kappos’ activities as principal investigator and member or chair of planning and steering committees or advisory boards in trials sponsored by Actelion, Addex, Bayer Health Care Pharmaceuticals, Bayer Schering Pharma, Biogen Idec, CLC Behring, Genentech, GeNeuro SA, Genzyme, Merck Serono, Mitsubishi Pharma, Novartis, Octapharma, Praxicon, Roche, Sanofi-Aventis, Santhera, Siemens and Teva; royalties from Neurostatus AG; research grants from the Swiss MS Society, Swiss National Research Foundation, the...
European Union, Gianni Rubatto Foundation, and the Novartis and Roche Research foundations.

**Funding**
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the University Hospital Basel and Novartis Pharma AG, Basel.

**References**


