Reference videos reduce variability of motor dysfunction assessments in multiple sclerosis

Marcus D'Souza, Saskia Steinheimer, Jonas Dorn, Cecily Morrison, Jacques Boisvert, Kristina Kravalis, Jessica Burggraaff ⬤, Caspar EP van Munster, Manuela Diederich, Abigail Sellen, Christian P Kamm, Frank Dahlke, Bernard MJ Uitdehaag and Ludwig Kappos

Abstract

Motor dysfunction, particularly ataxia, is one of the predominant clinical manifestations in patients with multiple sclerosis (MS). Assessment of motor dysfunction suffers from a high variability. We investigated whether the clinical rating of ataxia can be improved through the use of reference videos, covering the spectrum of severity degrees as defined in the Neurostatus-Expanded Disability Status Scale. Twenty-five neurologists participated. The variability of their assessments was significantly lower when reference videos were used (SD = 0.12; range = 0.40 vs SD = 0.26; range = 0.88 without reference videos; $p = 0.013$). Reference videos reduced the variability of clinical assessments and may be useful tools to improve the precision and consistency in the clinical assessment of motor functions in MS.

Keywords: Disability, video rating, ataxia, finger-to-nose test, EDSS

Introduction

In multiple sclerosis (MS) clinical assessment scales—mainly the Expanded Disability Status Scale (EDSS)—are used to quantify impairment and disability. The EDSS is known for a low inter- and intrarater reliability and suffers from a high variability, especially at lower EDSS scores.[1] Motor dysfunction and particularly ataxia is one of the predominant clinical manifestations in patients with MS and a major contributor to disability progression.[2] Thus, reliable and consistent rating of ataxia is crucial for the follow-up of patients with MS.

Objective

The objective of this report is to investigate whether reference videos (RVs) exemplifying degrees of ataxia severity can reduce the variability of motor dysfunction assessment in MS.

Methods

Study design and participants

This study was a subproject of "Assess MS,"[3] a study approved by the local ethics committees. All patients gave their written informed consent to the video recordings. Twenty-five raters (neurologists) from the university hospitals in Bern and Basel rated 60 videos based on 43 MS patients performing the finger-to-nose test (FNT). The videos were recorded with a Microsoft Kinect[TM] 1 camera and chosen out of >2000 videos recorded for the Assess MS study, with the constraint to have coverage for all limb ataxia grades of the Neurostatus-EDSS definitions.[4,5] According to these definitions there are five grades of limb ataxia: 0 = no ataxia, 1 = signs only, 2 = tremor or clumsy movements easily seen, minor interference with function, 3 = tremor or clumsy movements interfere with function in all spheres and 4 = most functions are very difficult. The ratings were performed at baseline and six weeks later ("retest"), to assess the long-term intrarater agreement. In both rating sessions 10% of the videos were presented twice for short-term intrarater agreement.

Forty-one RVs, different from the videos used for rating, were chosen by experienced neurologists of the Assess MS study. They also showed MS patients

Correspondence to:
Marcus D'Souza,
Neurologic Clinic and Policlinic, Departments of Medicine, Clinical Research, Biomedicine, and Biomedical Engineering, University Hospital of Basel, Petersgraben 4, CH-4031 Basel, Switzerland.
Marcus.Dsouza@usb.ch

Marcus D'Souza,
Neurologic Clinic and Policlinic, Departments of Medicine, Clinical Research, Biomedicine, and Biomedical Engineering, University Hospital of Basel, Switzerland

Saskia Steinheimer,
Department of Neurology, Inselspital, Bern University Hospital and University of Bern, Switzerland

Jonas Dorn,
Novartis Pharma AG, Switzerland

Cecily Morrison,
Microsoft Research, Cambridge, United Kingdom

Jacques Boisvert,
Novartis Pharma AG, Switzerland

Kristina Kravalis,
Neurologic Clinic and
Policlinic, Departments of
Medicine, Clinical Research,
Biomedicine, and
Biomedical Engineering,
University Hospital of Basel,
Switzerland

Jessica Burggraaff,
Caspar EP van Munster,
Department of Neurology,
VU University Medical
Center, The Netherlands

Manuela Diederich,
Neurologic Clinic and
Policlinic, Departments of
Medicine, Clinical Research,
Biomedicine, and
Biomedical Engineering,
University Hospital of Basel,
Switzerland

Abigail Sellen,
Novartis Pharma AG,
Switzerland

Christian P Kamm,
Department of Neurology,
Inselspital, Bern University
Hospital and University of
Bern, Switzerland
Neurology and
Neurorehabilitation Center,
Luzerner Kantonsspital,
Switzerland

Frank Dahlke,
Novartis Pharma AG,
Switzerland

Bernard MJ Uitdehaag,
Department of Neurology,
VU University Medical
Center, The Netherlands

Ludwig Kappos,
Neurologic Clinic and
Policlinic, Departments of
Medicine, Clinical Research,
Biomedicine, and
Biomedical Engineering,
University Hospital of Basel,
Switzerland

M.D.S. and S.S. contributed
equally to this work.

performing the FNT, with different degrees of limb-ataxia severity, based on the Neurostatus-EDSS definitions.[4,5] The raters were randomized into two groups: one group assessing videos based only on the written Neurostatus-EDSS definitions,[5] without simultaneous access to the RVs (Setting 1), and the other, with simultaneous access to the RVs (Setting 2). The characteristics of the raters are summarized in Table 1. There was no difference in experience with MS patients between the groups (Setting 1 vs Setting 2).

Patient performance

For FNT videos, MS patients were instructed by the recording neurologists of the Assess MS study to close their eyes and abduct their arms to 90 degrees at the shoulder in full extension, before touching the nose with the tip of their index finger, first with the dominant, then with the nondominant side (Figure 1).

Video rating

Videos were presented for rating on a touchscreen. Setting 2 allowed for simultaneous presentation of RVs on the right part of the screen (Figure 1). Horizontal swipe allowed for viewing RVs of different limb-ataxia severity degrees; vertical swipe for viewing alternative RVs of the same severity degree. In Setting 1 this part of the screen remained black. Raters were allowed to view each video as often as required for scoring.

Statistics

The analysis was conducted using Matlab R2014b (MathWorks, Natick, MA, USA). *F* test was used to compare the variability of the ratings between the two rater groups (Setting 1 vs Setting 2). Interrater agreement was calculated as intraclass correlation coefficient (ICC) for single measurements and absolute agreement.[6] Intrarater agreement was calculated as the percentage of identical ratings.

Results

The variability of ratings was significantly lower in Setting 2 (standard deviation (SD) = 0.12; range = 0.40) than in Setting 1 (SD = 0.26; range = 0.88, *F* test; $p = 0.013$), as illustrated in Figure 2. The ICC for interrater agreement was numerically slightly higher in Setting 2 (0.816 (95% confidence interval (CI): 0.756–0.871) vs 0.756 (95% CI: 0.674–0.829) in Setting 1) but this difference was not significant. Short-term and long-term intrarater agreement were similar across settings (Setting 1: 79±18% and 69 ±11%; Setting 2: 75±22% and 68±9%, not significant).

The average score of limb ataxia (according to the Neurostatus-EDSS definitions) was slightly higher in Setting 2, with RVs (mean score (test and retest after six weeks): 1.4 ± 0.1 in Setting 2, vs 1 ± 0.3 in Setting 1, $p < 0.0001$), as illustrated in Figure 2. No significant interaction was found between intrarater agreement, raters' experience with MS or EDSS assessments, or the centers (data not shown).

Discussion

As "pars pro toto," the results of this study show that using preselected RVs can reduce the rating variability in the assessment of limb ataxia of MS patients.

Table 1. Characteristics of patients and neurologists participating in this study.

Patients ($n = 43$)		
Age, years, mean ± SD (range)	42.79 ± 12.09 (23–77)	
Gender (female/male)	29/14	
Disease duration (years), mean ± SD (range)	13.25 ± 8.38 (0.5–40)	
Median EDSS (range)	3.5 (1–6.5)	
MS type, *n* (%)		
RRMS	39 (90.7%)	
SPMS	3 (7%)	
PPMS	1 (2.3%)	
Neurologists ($n = 25$)	Group Setting 1	Group Setting 2
Gender (female/male)	7/6	5/7
Years of experience with MS, mean (range)	5.5 (0.5–12)	5.8 (0.2–12)
Years of experience with EDSS, mean (range)	5.0 (0.5–12)	5.6 (0.2–12)
EDSS: Expanded Disability Status Scale; MS: multiple sclerosis; PPMS: primary progressive multiple sclerosis; RRMS: relapsing–remitting multiple sclerosis; SPMS: secondary progressive multiple sclerosis.		
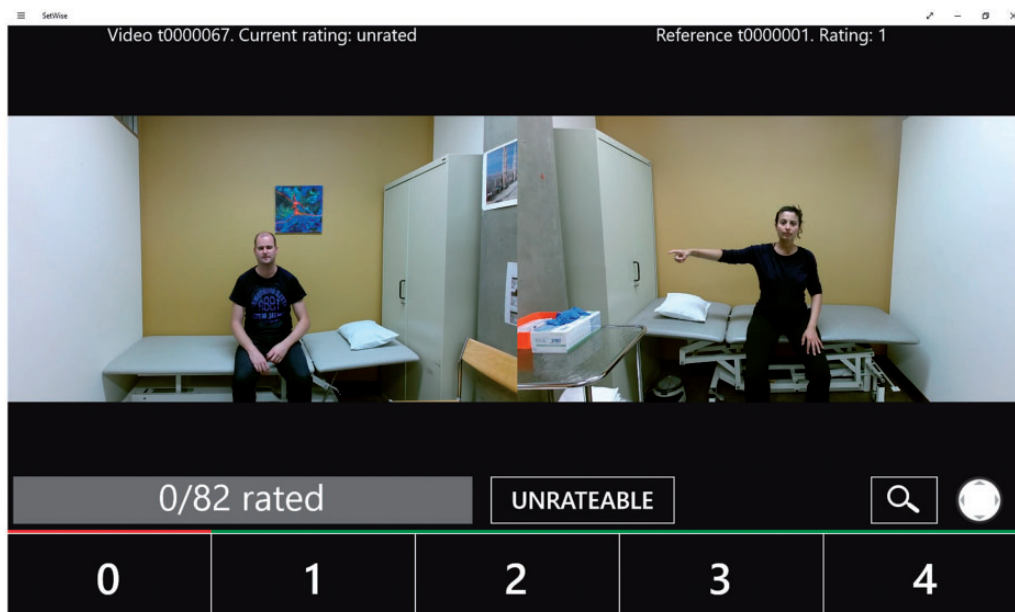
Figure 1. Reference videos on the right, videos to be rated on the left, below fields for scoring the appropriate severity of the performance using ataxia grades of the Neurostatus-Expanded Disability Status Scale definitions. According to these definitions there are five grades of limb ataxia: 0 = no ataxia, 1 = signs only, 2 = tremor or clumsy movements easily seen, minor interference with function, 3 = tremor or clumsy movements interfere with function in all spheres and 4 = most functions are very difficult. People shown are not patients and gave written consent to be shown.
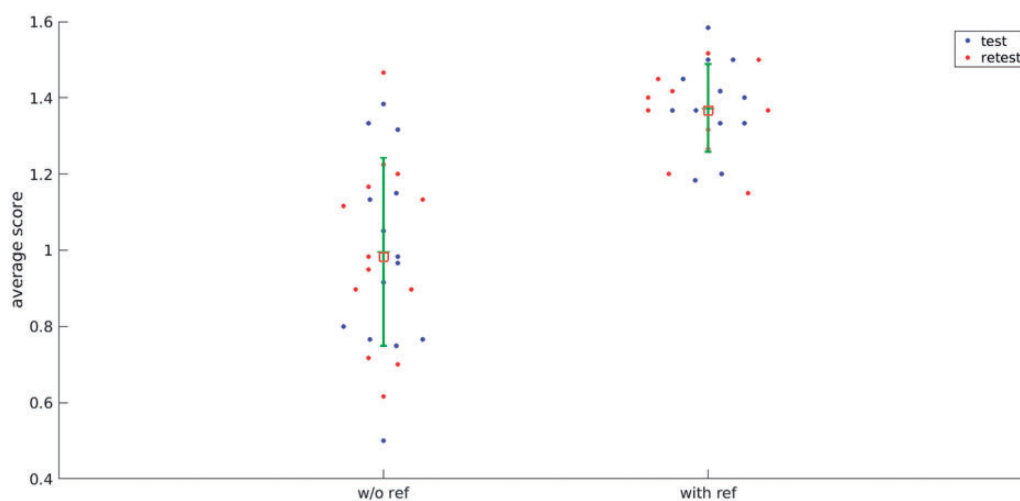
Figure 2. On the left the ratings of Setting 1, i.e. the group without reference videos are shown without ((w/o) ref) and on the right, those from Setting 2, i.e. the group with reference videos ("with ref"). Mean and standard deviation (SD) are shown in green, median in magenta. The variability of ratings was significantly lower in Setting 2 (SD = 0.12; range = 0.40) than in Setting 1, w/o) reference videos (SD = 0.26; range = 0.88, F test; $p = 0.013$). Each dot represents the average of all ratings of one neurologist (blue at baseline and red six weeks later).

The use of such videos can be easily implemented and does not require an additional/new scale, since we used the already well-established Neurostatus-EDSS definitions.[4] Whether this approach can also be used for assessments other than limb ataxia remains to be shown.

We found a small but statistically significant difference of the average severity level obtained in the two settings with higher ratings in the setting with RVs. As the ataxia degrees were assigned to the RVs by neurologists with special expertise in clinical ratings, this may have contributed to stricter interpretation of

the grade definitions. A further limitation in our study was the low number of severely affected patients (ataxia grades 3 and 4). In daily routine, however, rating of lower-severity grades is more challenging than higher grades. Our RV approach may also have a role in training machine-learning algorithms (MLAs). Such an example is the Assess MS system, a potentially finer-grained tool to measure motor dysfunction in MS.[3] This system uses advanced MLAs to analyze three-dimensional-depth-sensor recordings of MS patients performing standard tests of motor function, like the FNT. Reducing the variability of clinical assessments that are used to train MLAs should also contribute to improved algorithms that are derived from machine learning.

Conclusions

The use of RVs may represent a simple method to reduce variability in the assessment of motor dysfunction in MS. This method could be particularly useful in the context of clinical research, and to train MLAs.

Conflict of Interests

Funding

Supplemental material

Supplemental material is available for this article online.

ORCID iD

Jessica Burggraaff http://orcid.org/0000-0002-9367-9686

References

1. Meyer-Moock S, Feng YS, Maeurer M, et al. Systematic literature review and validity evaluation of the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) in patients with multiple sclerosis. *BMC Neurol* 2014; 14: 1–10.

2. Weinshenker BG, Issa M and Baskerville J. Long-term and short-term outcome of multiple sclerosis: A 3-year follow-up study. *Arch Neurol* 1996; 53: 353–358.

3. Morrison C, D'Souza M, Huckvale K, et al. Usability and acceptability of ASSESS MS: Assessment of motor dysfunction in multiple sclerosis using depth-sensing computer vision. *JMIR Hum Factors* 2015; 2: e11.

4. D'Souza M, Yaldizli Ö, John R, et al. Neurostatus e-Scoring improves consistency of Expanded Disability Status Scale assessments: A proof of concept study. *Mult Scler* 2017; 23: 597–603.

5. Kappos L. Neurostatus scoring definitions (version 04/10.2), 2011, http://www.neurostatus.net.

6. McGraw KO and Wong SP. Forming inferences about some intraclass correlation coefficients. *Psych Meth* 1996; 1: 30–46.