Clinical Neurophysiology 130 (2019) 307–314

Contents lists available at ScienceDirect

Clinical Neurophysiology

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

Diagnostic criteria for amyotrophic lateral sclerosis: A multicentre study of inter-rater variation and sensitivity

B. Johnsen [a,*], K. Pugdahl [a], A. Fuglsang-Frederiksen [a], K. Kollewe [b], L. Paracka [b], R. Dengler [b], J.P. Camdessanché [c], W. Nix [d], R. Liguori [e], I. Schofield [f], L. Maderna [g], D. Czell [h], C. Neuwirth [i], M. Weber [i], V.E. Drory [j], A. Abraham [j], M. Swash [k], M. de Carvalho [l]

[a] Department of Clinical Neurophysiology, Aarhus University Hospital, Aarhus, Denmark
[b] Department of Neurology, Hannover Medical School, Hannover, Germany
[c] Department of Neurology, Saint-Etienne University Hospital, Saint-Etienne, France
[d] Department of Neurology, University Clinics Mainz, Mainz, Germany
[e] IRCCS Institute of Neurological Sciences of Bologna and Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy
[f] Department of Clinical Neurophysiology, Newcastle General Hospital, Newcastle upon Tyne, UK
[g] Department of Neurology and Laboratory of Neuroscience, IRCCS Istituto Auxologico Italiano, University of Milano, Milan, Italy
[h] Neurology, Spital Linth, Uznach, Switzerland
[i] Neuromuscular Diseases Unit/ALS Clinic, Cantonal Hospital St. Gallen, St. Gallen, Switzerland
[j] Neuromuscular Service of the Department of Neurology, Tel Aviv Sourasky Medical Center, Tel Aviv University, Tel Aviv, Israel
[k] Barts and the London School of Medicine, Queen Mary University of London, London, UK
[l] Institute of Physiology-Instituto de Medicina Molecular, Faculty of Medicine, University of Lisbon, and Department of Neurosciences, Hospital de Santa Maria-CHLN, Lisbon, Portugal

See Editorial, pages 303–304

ARTICLE INFO

Article history:
Accepted 11 November 2018
Available online 9 December 2018

Keywords:
Amyotrophic lateral sclerosis
Awaji criteria
Revised El Escorial criteria
Diagnostic criteria
Inter-rater variation
Electromyography
Electrodiagnosis

HIGHLIGHTS

- Inter-rater variation in the use of diagnostic criteria for amyotrophic lateral sclerosis is high.
- Revised El Escorial and Awaji criteria are complex and might require training for use.
- The Awaji criteria have a positive impact on diagnostic sensitivity in subgroups.

ABSTRACT

Objective: This study assesses inter-rater agreement and sensitivity of diagnostic criteria for amyotrophic lateral sclerosis (ALS).
Methods: Clinical and electrophysiological data of 399 patients with suspected ALS were collected by eleven experienced physicians from ten different countries. Eight physicians classified patients independently and blinded according to the revised El Escorial Criteria (rEEC) and to the Awaji Criteria (AC). Inter-rater agreement was assessed by Kappa coefficients, sensitivity by majority diagnosis on 350 patients with follow-up data.
Results: Inter-rater agreement was generally low both for rEEC and AC. Agreement was best on the categories "Not-ALS", "Definite", and "Probable", and poorest for "Possible" and "Probable Laboratory-supported".

Sensitivity was equal for rEEC (64%) and AC (63%), probably due to downgrading of "Probable Laboratory-supported" patients by AC. However, AC was significantly more effective in classifying patients as "ALS" versus "Not-ALS" ($p < 0.0001$).
Conclusions: Inter-rater variation is high both for rEEC and for AC probably due to a high complexity of the rEEC inherent in the AC.

The gain of AC on diagnostic sensitivity is reduced by the omission of the "Probable Laboratory-supported" category.

* Corresponding author at: Department of Clinical Neurophysiology, Aarhus University Hospital, 8000 Aarhus C, Denmark.
E-mail address: birgjohn@rm.dk (B. Johnsen).

https://doi.org/10.1016/j.clinph.2018.11.021

Significance: The results highlight a need for initiatives to develop simpler and more reproducible diagnostic criteria for ALS in clinical practice and research.

1. Introduction

Diagnostic criteria for amyotrophic lateral sclerosis (ALS) have been developed to standardise diagnosis and patient recruitment in trials. A research group of the World Federation of Neurology (WFN) devised the "El Escorial criteria" (EEC) (Brooks et al., 1994), which were later modified to increase their diagnostic sensitivity as the "revised El Escorial Criteria" (rEEC) (Brooks et al., 2000) with the objective of encouraging earlier inclusion in trials (Ross et al., 1998). The rEEC have been found to have high specificity, but to lack sensitivity, thus failing to promote earlier patient inclusion in clinical trials (Traynor et al., 2000; Zoccolella et al., 2006; Turner et al., 2009). In order to increase the sensitivity of the diagnostic criteria, an amendment to the rEEC was suggested by a group sponsored by the International Federation of Clinical Neurophysiology (IFCN) at a meeting at Awaji-shima, Japan (de Carvalho et al., 2008). These "Awaji criteria" (AC) refined the applied electrodiagnostic algorithm for detection of lower motor neuron (LMN) involvement to support ALS diagnosis by incorporating a number of new concepts. The major differences, compared to the rEEC, were that: (1) clinical and electrophysiological findings of LMN involvement should have equal significance to determine involvement of a specific region; (2) the category "Probable Laboratory-supported" was rendered redundant; and (3) the presence of fasciculation potentials in muscles with evidence of re-innervation should serve as evidence of ongoing denervation, analogous to the presence of fibrillation potentials and positive sharp-waves; a feature particularly relevant in cranial-innervated muscles and in muscles with preserved strength.

The diagnostic sensitivities of rEEC and AC have been compared in three meta-analyses (Costa et al., 2012; Jang and Bae, 2015; Geevasinga et al., 2016b), the latter studying individual patient data from all previous studies. One meta-analysis did not favour AC (Jang and Bae, 2015), but in the two others (Costa et al., 2012; Geevasinga et al., 2016b), AC demonstrated a higher sensitivity with similar specificity. AC was particularly favourable in bulbar-onset patients (de Carvalho and Swash, 2009; Okita et al., 2011) and in patients with a short-disease duration (between 6–11 months) before investigation (Geevasinga et al., 2016b). The same advantages for AC were found in the single multicentre-study so far published comparing rEEC and AC (Geevasinga et al., 2016a), in which AC gained 12% in sensitivity, with the same high specificity (99.5%).

In some other studies the positive impact of AC was modest but increased by the re-inclusion of the "Probable Laboratory-supported" category in the AC (Boekestein et al., 2010). The positive impact of this category was particularly relevant in patients with limb-onset disease (Geevasinga et al., 2016b). Originally, both for rEEC and AC, the intention was to include only "Definite" and "Probable" cases in clinical trials, but recently it has been shown that additional inclusion of "Possible" cases may increase sensitivity without lowering the specificity (Geevasinga et al., 2016a).

The rEEC and the AC were derived by consensus as research criteria, not as criteria for diagnosis in clinical practice. There are, however, no formal diagnostic criteria for ALS intended for use in ordinary clinical practice, and probably due to this lack some may use the rEEC or the AC for clinical practice.

It is essential that diagnostic criteria for ALS show a high specificity and sensitivity, and especially that they are reproducible between raters. In previous studies on sensitivity and specificity, the reproducibility of the criteria specified has only scarcely been addressed. Forbes et al. (2001) found good reproducibility regarding ALS versus non-ALS diagnosis between two raters' classifications of 65 patients according to the EEC and rEEC, without significant differences between the two sets of criteria. The aim of this study was to evaluate inter-rater reliability of rEEC and AC among raters without previous formal training regarding use of these classificatory systems. We also compared the sensitivity and specificity of rEEC and AC among these raters.

2. Methods

2.1. Inclusion

The study was carried out as an international multicentre initiative involving physicians from 11 different centres in Europe and Israel, prompted by the European multicentre collaboration ESTEEM (Vingtoft et al., 1995). The participating physicians were recruited based on their experience in neurophysiology and interest in ALS.

This study included consecutive patients with a clinical suspicion of ALS, referred for electrophysiological investigation at the participating centres, in order to confirm or refute the clinical diagnosis. Patients with any unequivocal alternative diagnosis were excluded.

Three-hundred-ninety-nine patients (224 men; 163 women; 12 not registered) with a mean age of 63.6 years (SD 12.2) were included. Two-hundred and ten patients (53%) had spinal-onset, 105 (26%) had bulbar-onset, 9 (2.3%) had axial-onset, and 8 (2.0%) had generalised onset. The region of onset was not established in 67 patients (17%). Mean disease duration was 16.3 months (range 2–168 months). Disease duration was less than 12 months for 185 patients, 12 months or more for 208 patients and unknown for 6 patients.

At least three regions should be tested by concentric needle EMG as required by the Awaji neurophysiological algorithm. In the cervical and lumbosacral regions at least two muscles innervated by different roots and nerves should be sampled, while at least one muscle should be investigated in the bulbar or thoracic region (de Carvalho et al., 2008). For each tested muscle, force on the MRC scale and atrophy (0, +, ++, +++) were recorded. In addition, at least three motor nerves, including at least one in a wasted limb, and at least one sensory nerve in a wasted limb should be tested.

Patient data were anonymised, and each patient assigned a number. The participating centres obtained permission according to local ethical requirements.

2.2. Data structure

2.2.1. Clinical data

Clinical data were reported in a standardised format (example in Supplementary Data). For each of the four regions the presence of upper motor neuron (UMN) signs (abnormally brisk or irradiating deep tendon reflexes, spasticity, Hoffman sign, brisk jaw jerk, and extensor plantar reflex) and LMN signs (weakness, atrophy, fasciculations, and hyporeflexia) were entered. In addition, age, sex, disease duration, progressive disease, onset region, ALS-FRS and ALS-FRS bulbar scores, forced vital capacity, and results of diagnostic tests, especially imaging, as well as medication were noted.

2.2.2. Electrodiagnostic data

EMG data were presented in predefined forms allowing for flexible entry of the test parameters by the examining centres.

For spontaneous activity, presence of fibrillations, positive sharp waves, and fasciculations were noted (0, +, ++, +++, ++++) together with electrophysiological stability of fasciculations (stable/unstable). For the interference pattern at maximal voluntary contraction, pattern (full/reduced/discrete/single units/ submaximal/no recruitment), recruitment rate (normal/increased) and amplitude (normal/increased/decreased/borderline increased or decreased) were noted, while for motor unit potential (MUP) analysis duration (normal/increased/decreased/borderline increased or decreased), amplitude, number of polyphasic potentials (normal/increased/increased borderline), and stability were noted.

For nerves, motor or sensory conduction velocity, compound muscle action potential or sensory nerve action potential amplitude, F-wave persistence were noted as normal/decreased/decreased borderline, while distal motor latency and F-wave latency were noted as normal/increased/increased borderline. In addition, presence of conduction block or absence of response was noted.

2.2.3. Diagnoses

The physicians' classifications according to the rEEC ("Definite", "Probable", "Probable Laboratory-supported", "Possible", "Not-ALS") and the AC ("Definite", "Probable", "Possible", "Not-ALS") were noted on the data sheets.

2.3. Data collection procedure

The data collection procedure is outlined in Fig. 1. Clinical and electrophysiological data from patients examined according to the local practice at the submitting centres were entered on standardised data sheets.

The data sheets were sent electronically to the coordinating centre (Aarhus University Hospital, Denmark) for data validation. A total of 465 examination data sheets were submitted from 12 different centres. Of these, 66 were discarded as they did not fulfil the inclusion criteria, mainly due to EMG performed in two regions only. Of the examinations used for further analyses, 11 centres submitted 7, 7, 11, 12, 17, 28, 37, 44, 62, 70, and 104 examinations.

The 399 examinations fulfilling the inclusion criteria were then reformatted by removing the examiner's affiliation and diagnostic classifications and sent to all 11 centres for independent blind evaluation with respect to classification according to rEEC and AC. All the data sheets were returned to Aarhus by eight centres, which gave a total of 3192 classifications for each of the criteria sets.

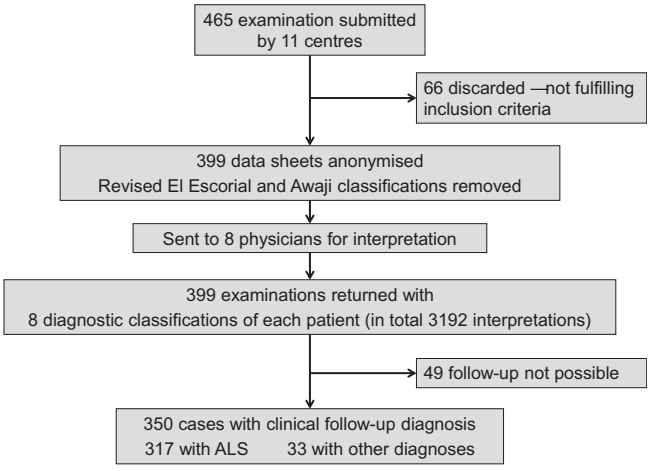

Fig. 1. Flowchart of study on inter-rater variation of diagnostic criteria for ALS.

A clinical follow-up with respect to final diagnosis, with respect to ALS or not, was requested for all 399 patients. Follow-up information was received for 350 patients as 49 patients were lost to follow-up.

2.4. Majority diagnoses

A majority diagnosis for each patient was decided as the category chosen most often by the eight assessing physicians according to rEEC and AC in order to assess sensitivity and specificity. If there was the same number of classifications for two categories, the category with the highest probability for ALS was arbitrarily chosen, e.g. if there were four classifications of both "Probable" and "Possible", "Probable" was chosen.

2.5. Statistical analysis

Agreement among raters was assessed by the Kappa coefficient for multiple raters (Fleiss, 1971). Confidence intervals for Kappa coefficients were calculated as described by Zou and Donner (2004). Comparison of two Kappa coefficients was performed by assessment of the 95% confidence interval for their difference based on bootstrap with 5000 repetitions.

Sensitivities and specificities for rEEC and AC were compared with McNemar's tests, and sensitivities and specificities between different groups of onset region and disease duration were compared by Chi-squared tests.

Stata version 15 with the additional package "kaputil" and Microsoft Excel were used for statistical calculations.

Statistical significance was defined as 5%.

3. Results

3.1. Distribution of classifications

The distribution of the 3192 classifications in the diagnostic categories of rEEC and AC are shown in Fig. 2. There were more classifications as "Definite" (681 vs. 343; Chi2 132.9; $p < 0.00001$) and "Probable" (1167 vs. 943; Chi2 35.5; $p < 0.00001$) using the AC compared with the rEEC. However, when including the "Probable Laboratory-supported" category of rEEC, there were fewer classifications as "Probable" using the AC than the rEEC (1167 vs. 1434; Chi2 46.3; $p < 0.00001$). Looking at the pooled number of classifications "Definite/Probable/Probable Laboratory-supported" there was a non-significant advantage for AC (1848 vs. 1777; Chi2 3.2; $p = 0.07$). There was a non-significant difference of more cases classified as "Possible" by AC than by rEEC (781 vs. 717; Chi2 3.6; $p = 0.06$). AC was significantly more effective in classifying patients as "ALS" versus "Not-ALS" (698 "Not-ALS" with rEEC vs. 563 "Not-ALS" with AC; Chi2 18.0; $p < 0.0001$).

3.2. Inter-rater variation on categories

There was a rather large variation among the eight physicians in the number of cases classified in the different categories both for rEEC and AC (Table 1). The dichotomy "Definite/Probable" versus "Possible/Not-ALS", reflecting the number of patients possibly included in a trial, ranged from 142 to 308 patients (35.6–77.2%) according to the rEEC, and from 157 to 299 patients (39.3–74.9%) according to the AC. Considering the dichotomy "Definite/Probable/Possible" versus "Not-ALS", the variation among the eight physicians was smaller, from 241 (60.4%) to 338 (84.7%) according to the rEEC and from 286 (71.7%) to 345 (86.5%) according to the AC.

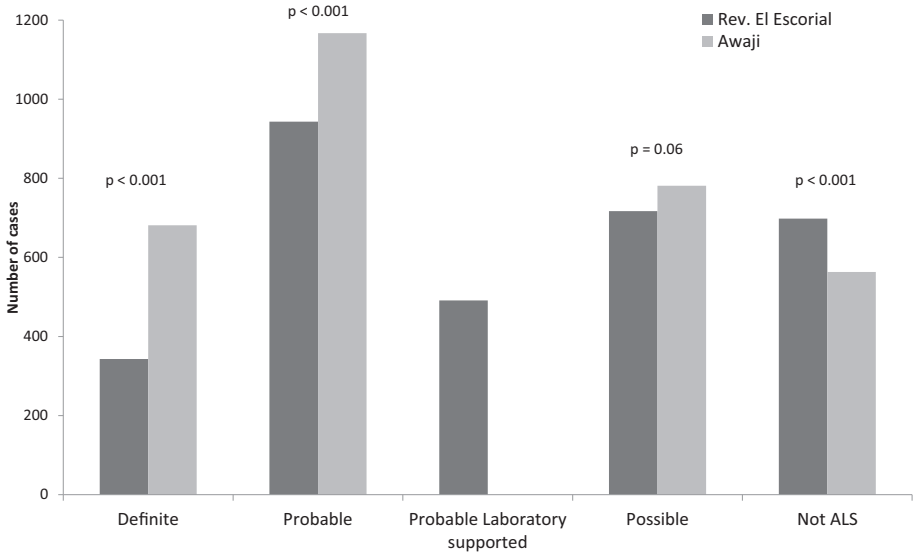

Fig. 2. Eight physicians' classifications of 399 cases referred for ALS according to the revised El Escorial and the Awaji criteria. A total of 3192 classifications were distributed in five categories for the revised El Escorial criteria and in four categories for the Awaji criteria. Differences between revised El Escorial and Awaji criteria were calculated by Chi-squared tests (p-values shown).

Table 1
Diagnostic classifications by eight physicians.

Rev. El Escorial	Definite	Probable	Prob. Lab. sup.	Possible	Not ALS	Def/Prob	Def/Prob/Poss
	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
Physician #1	58 (1,5)	136 (3,4)	49 (1,2)	95 (2,4)	61 (1,5)	243 (6,1)	338 (8,5)
Physician #2	50 (1,3)	137 (3,4)	81 (2,0)	70 (1,8)	61 (1,5)	268 (6,7)	338 (8,5)
Physician #3	50 (1,3)	140 (3,5)	108 (2,7)	38 (1,0)	63 (1,6)	298 (7,5)	336 (8,4)
Physician #4	58 (1,5)	127 (3,2)	123 (3,1)	20 (0,5)	71 (1,8)	308 (7,7)	328 (8,2)
Physician #5	17 (0,4)	93 (2,3)	60 (1,5)	155 (3,9)	74 (1,9)	170 (4,3)	325 (8,1)
Physician #6	41 (1,0)	119 (3,0)	17 (0,4)	137 (3,4)	85 (2,1)	177 (4,4)	314 (7,9)
Physician #7	46 (1,2)	96 (2,4)	29 (0,7)	103 (2,6)	125 (3,1)	171 (4,3)	274 (6,9)
Physician #8	23 (0,6)	95 (2,4)	24 (0,6)	99 (2,5)	158 (4,0)	142 (3,6)	241 (6,0)
minimum	17 (0,4)	93 (2,3)	17 (0,4)	20 (0,5)	61 (1,5)	142 (3,6)	241 (6,0)
maximum	58 (1,5)	140 (3,5)	123 (3,1)	155 (3,9)	158 (4,0)	308 (7,7)	338 (8,5)
Awaji	Definite	Probable	Prob. Lab. sup.	Possible	Not ALS	Def/Prob	Def/Prob/Poss
Physician #1	88 (2,2)	156 (3,9)		101 (2,5)	54 (1,4)	244 (6,1)	345 (8,6)
Physician #2	70 (1,8)	224 (5,6)		47 (1,2)	58 (1,5)	294 (7,4)	341 (8,5)
Physician #3	92 (2,3)	157 (3,9)		93 (2,3)	57 (1,4)	249 (6,2)	342 (8,6)
Physician #4	196 (4,9)	103 (2,6)		39 (1,0)	61 (1,5)	299 (7,5)	338 (8,5)
Physician #5	36 (0,9)	121 (3,0)		170 (4,3)	72 (1,8)	157 (3,9)	327 (8,2)
Physician #6	56 (1,4)	140 (3,5)		134 (3,4)	69 (1,7)	196 (4,9)	330 (8,3)
Physician #7	95 (2,4)	133 (3,3)		92 (2,3)	79 (2,0)	228 (5,7)	320 (8,0)
Physician #8	48 (1,2)	133 (3,3)		105 (2,6)	113 (2,8)	181 (4,5)	286 (7,2)
minimum	36 (0,9)	103 (2,6)		39 (1,0)	54 (1,4)	157 (3,9)	286 (7,2)
maximum	196 (4,9)	224 (5,6)		170 (4,3)	113 (2,8)	299 (7,5)	345 (8,6)

Individual classifications according to the revised El Escorial criteria and to the Awaji criteria by eight physicians on 399 cases referred for ALS. Prob. Lab. sup., Probable Laboratory-supported; Def/Prob, Sum of Definite and Probable; Def/Prob/Poss, Sum of Definite, Probable, and Possible.

3.3. Agreement on classifications

Kappa coefficients for the agreement among the eight physicians on the diagnostic categories are shown in Fig. 3. Both for rEEC and for AC, the agreement was best in the "Not-ALS" category with Kappa coefficients of 0.59 (95% CI:0.52–0.65) for rEEC and 0.65 (95% CI:0.58–0.72) for AC. Both for rEEC and AC, the agreement in the "Definite" and the "Probable" categories were lower than in the "Not-ALS" category (Fig. 3), and agreement was lowest in the "Possible" category with Kappa coefficients of 0.14 (95% CI:0.09–0.20) for rEEC and 0.33 (95% CI:0.26–0.40) for AC. For the rEEC category "Probable Laboratory-supported" there was also a low level of agreement with a Kappa coefficient of 0.25 (95% CI:0.18–0.34). For the pooled rEEC category of "Probable" and "Probable Laboratory-supported", the Kappa coefficient was 0.36 (95% CI:0.32–0.41).

3.4. Differences in agreement between the two sets of criteria

Both for the rEEC and AC, agreement among the raters varied in the different diagnostic categories (Fig. 3). Agreement was higher for the AC than for the rEEC in the "Not-ALS" category (95% CI for difference: 0.03–0.10; $p < 0.001$, bootstrap) and in the "Possible" category (95% CI for difference: 0.15–0.23; $p < 0.001$, bootstrap). In contrast, agreement was higher for the rEEC than for the AC in the "Definite" category (95% CI for difference: 0.00–0.12; $p = 0.04$, bootstrap) and in the "Probable" category (95% CI for difference: 0.11–0.21; $p < 0.001$, bootstrap). The Kappa coefficient of the pooled rEEC category of "Probable" and "Probable Laboratory-supported" did not differ from that of the "Probable" category of AC (95% CI for difference: −0.06–0.03; $p = 0.45$, bootstrap).

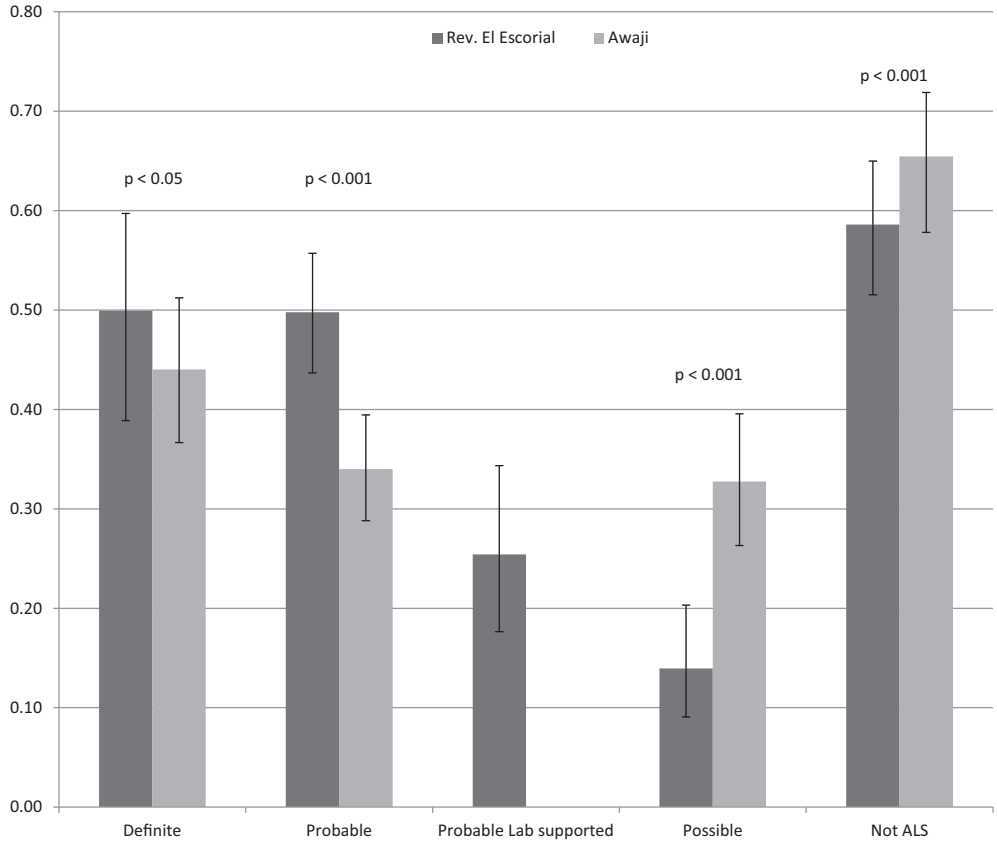

Fig. 3. Inter-rater agreement among eight experienced physicians on diagnostic categories of the revised El Escorial criteria and the Awaji criteria on 399 patients referred for ALS assessed by Kappa coefficients. Differences between agreement on the revised El Escorial and the Awaji criteria were calculated by a bootstrap method (p-values shown).

3.5. Sensitivity and specificity

The majority diagnoses were used to calculate sensitivities and specificities for rEEC and AC for all 350 patients with available follow-up data (Table 2). There were no differences between the sensitivities or specificities of rEEC and AC for neither all patients nor subgroups with limb onset or disease duration < 12 months (McNemar's' tests; $p > 0.05$). When "Possible" patients were included, however, there was a higher sensitivity of AC than rEEC for the subgroups of bulbar onset patients and patients with disease duration $\geq$ 12 months (McNemar's' tests, $p = 0.02$ and $p = 0.03$, respectively) (Table 2).

Both for rEEC and AC, there were no differences in sensitivity or specificity between the groups of patients with bulbar-onset and limb-onset, or between patients with disease duration < 12 months and $\geq$ 12 months (Chi-squared, $p > 0.05$).

3.6. Changes from revised El Escorial to Awaji

The distribution of corresponding AC classifications for each of the five rEEC categories are shown in Fig. 4, based on the majority diagnoses for the 317 patients in whom a diagnosis of ALS was confirmed at clinical follow-up.

For 45 patients classified as "Definite" by the rEEC, all but one was also classified as "Definite" by the AC. For 122 patients classified as "Probable" by the rEEC most (99 [81.1%]) were also classified as "Probable" by the AC, but 20 (16.4%) of these were classified by AC as "Definite". For 36 patients classified as "Probable Laboratory-supported" only 11 (30.6%) were classified as "Probable" by the AC. Most of these (20 [55.6%]) were classified with a lower probability as "Possible", and 5 (13.9%) with a higher probability as "Definite". For 69 patients classified as "Possible" by the rEEC, 49 (71.0%) were also classified as "Possible" by the AC; but 15 (21.7%) were reclassified as "Probable". Of the 45 patients who were classified as "Not-ALS" by the rEEC, 31 (68.9%) remained as "Not-ALS" by the AC, but 14 patients were classified with a higher probability as "Possible" (10 patients [22.2%]) or "Probable" (4 patients [8.9%]) by the AC.

For all 317 patients, 57 (18.0%) were classified with a higher probability by the AC than by the rEEC, and 26 (8.2%) were classified with a lower probability. Of the 26 patients classified with a lower probability, 20 shifted from rEEC "Probable Laboratory-supported" to "Possible".

The number of patients to be included in clinical trials was 203 (64.0%) for the rEEC and 201 (63.4%) for the AC when using the classical criteria of including only patients with "Definite" and "Probable" disease. Extending the group of patients to be included with the "Possible" category, resulted in 272 (85.8%) to include by using the rEEC and 281 (88.6%) to include by the AC.

4. Discussion

4.1. Study design

To compare different sets of diagnostic criteria, a multicentre design is the most appropriate approach. For diagnosis of ALS only the study by Geevasinga et al. (2016a) has used this design focusing on diagnostic sensitivities and specificities without considering inter-rater variation.

In our study, we added a number of new challenges, we recruited patients from 11 different centres in 10 countries, the neurophysiologists in these centres did not receive any special training, and their neurophysiological interpretations were not standardized. Furthermore, we included analyses of inter-rater agreement using reviews

Table 2
Sensitivity and specificity of diagnostic criteria for ALS.

	El Escorial (95% CI)	Awaji (95% CI)	McNemar
All patients (n = 350)			
Sensitivity (def/prob)	0.64 (0.58–0.69)	0.63 (0.58–0.69)	NS
Sensitivity (def/prob/poss)	0.86 (0.81–0.89)	0.89 (0.85–0.92)	NS
Specificity (def/prob)	0.91 (0.74–0.98)	0.85 (0.67–0.94)	NS
Specificity (def/prob/poss)	0.82 (0.64–0.92)	0.76 (0.76–0.88)	NS
Bulbar onset (n = 97)			
Sensitivity (def/prob)	0.63 (0.53–0.73)	0.69 (0.58–0.78)	NS
Sensitivity (def/prob/poss)	0.83 (0.73–0.90)	0.92 (0.85–0.97)	p = 0.016
Specificity (def/prob)	0.75 (0.22–0.99)	0.75 (0.22–0.99)	NS
Specificity (def/prob/poss)	0.75 (0.22–0.99)	0.75 (0.22–0.99)	NS
Limb onset (n = 186)			
Sensitivity (def/prob)	0.59 (0.51–0.66)	0.57 (0.49–0.64)	NS
Sensitivity (def/prob/poss)	0.85 (0.79–0.90)	0.85 (0.79–0.90)	NS
Specificity (def/prob)	0.91 (0.69–0.98)	0.86 (0.64–0.96)	NS
Specificity (def/prob/poss)	0.82 (0.59–0.94)	0.85 (0.79–0.90)	NS
Disease duration < 12 months (n = 159)			
Sensitivity (def/prob)	0.64 (0.56–0.72)	0.66 (0.58–0.73)	NS
Sensitivity (def/prob/poss)	0.89 (0.82–0.93)	0.89 (0.83–0.94)	NS
Specificity (def/prob)	1.00 (0.56–1.00)	0.86 (0.42–0.99)	NS
Specificity (def/prob/poss)	0.86 (0.42–0.99)	0.71 (0.30–0.95)	NS
Disease duration >= 12 months (n = 185)			
Sensitivity (def/prob)	0.64 (0.56–0.71)	0.61 (0.53–0.69)	NS
Sensitivity (def/prob/poss)	0.83 (0.76–0.88)	0.88 (0.82–0.93)	p = 0.027
Specificity (def/prob)	0.88 (0.68–0.97)	0.84 (0.63–0.95)	NS
Specificity (def/prob/poss)	0.80 (0.59–0.92)	0.76 (0.54–0.90)	NS

Sensitivity and specificity of revised El Escorial criteria and Awaji criteria on 350 patients referred for ALS. Data for all patients and for subgroups according to onset region and disease duration are shown. Diagnoses are majority diagnoses of eight experienced physicians on 317 patients with a follow-up diagnosis of ALS and 33 patients where the ALS diagnosis could not be confirmed at follow-up. CI, confidence interval; NS, not significant; def, definite; prob, probable; poss, possible.

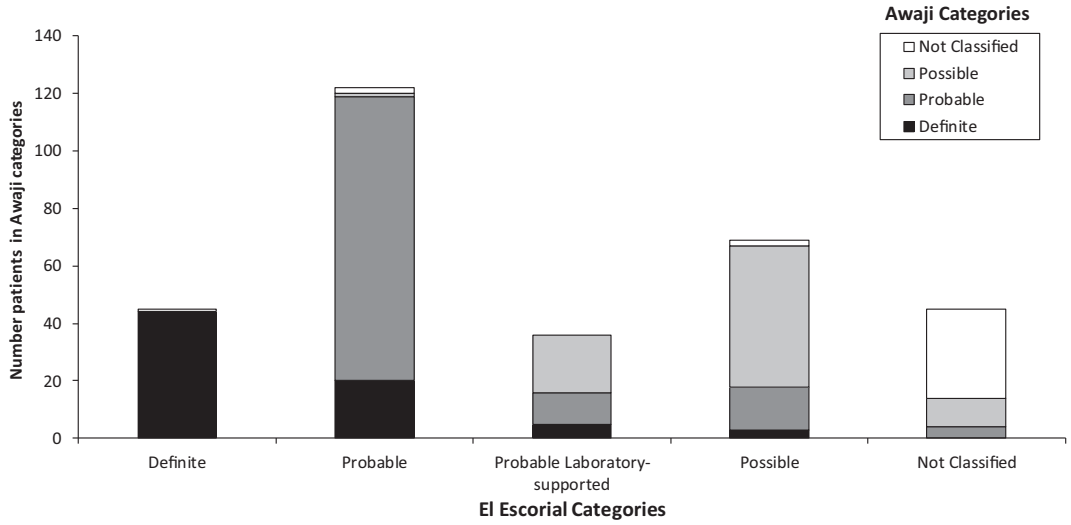

Fig. 4. Changes from El Escorial Categories to Awaji Categories. Distribution of classifications according to the Awaji criteria for each of the five categories of the revised El Escorial criteria. Diagnoses are majority diagnoses of eight experienced physicians on 317 patients with a follow-up diagnosis of ALS.

of the data sheets by 8 of the participating neurophysiologists. In considering the results of our study it should be borne in mind that there is no specific one-off diagnostic test for ALS, and that the diagnostic standard against which any set of diagnostic criteria must be tested is a clinical opinion, made at the time of clinical assessment, but supplemented for diagnostic certainty by progression of the disease over time. Any set of definitions for use at a single time-point, although necessary for clinical trials, is inevitably limited by the conflicting requirements of specificity and sensitivity.

4.2. Inter-rater agreement

We found a rather large variation among the included neurophysiologists. The best agreement was on the important "ALS/Not-ALS" dichotomy, in particular when using AC. Substantial disagreement was observed on the "Possible" category for both rEEC and AC, but the agreement was better for AC (Fig. 3). The "Probable Laboratory-supported" category of the rEEC disclosed a very poor inter-rater agreement, probably because it requires a combination of clinical and electrodiagnostic interpretation. Agreement was lowest in the "Possible" and "Probable Laboratory-supported" categories. This may because these categories represent borderline cases where most disagreement is to be expected. There was only a minor difference in the agreement level between AC and rEEC with respect to the "Definite" category. In the "Probable" category, there was no difference in the agreement between AC and rEEC, when the rEEC categories of "Probable" and "Probable Laboratory-supported" were pooled.

The rather low levels of agreement resulted in a large inter-rater variation in number of patients to be included in trials, most pronounced if patients with only "Definite" and "Probable" ALS were included, and less pronounced if also patients with "Possible" ALS were included (Table 1). Using the latter approach of including "Possible", the maximal difference between raters in number of patients to include was lower using AC (59 patients) than using rEEC (97 patients) (Table 1), indicating that using AC for including patients with "Definite /Probable/Possible" disease in a trial would provide a more homogenous population across different centres. It was striking that 66 of the original 465 patients were excluded mainly because their neurophysiological examinations did not test the required 3 regions necessary for the Awaji criteria to be fulfilled. This indicates a degree of diagnostic confidence that should perhaps be considered in any future revision of the current diagnostic criteria.

4.3. Comparison with other studies of inter-rater agreement

Forbes (2001) found a Kappa coefficient of 0.88 (95% CI: 0.80–0.96) for the "ALS/Not-ALS" dichotomy for the rEEC. Although Kappa coefficients are not necessarily comparable across studies (Maclure and Willett, 1987), this value is higher than ours of 0.59 (95% CI: 0.52–0.65). The most likely explanation for this is the larger number of contributing centres and raters in our study, resulting in greater variance.

4.4. Sensitivity and specificity

The diagnostic sensitivity of rEEC was 0.64 (95% CI: 0.58–0.69) in our study, a result comparable with that of the largest meta-analysis of 0.62 (95% CI: 0.49–0.75) (Costa et al., 2012). For AC the diagnostic sensitivity was 0.63 (95% CI: 0.58–0.69), compared with that of the meta-analysis of 0.81 (95% CI: 0.72–0.90). We could not confirm, therefore, that AC significantly increases the sensitivity compared with rEEC (Costa et al., 2012). However, when "Possible" patients are included AC was more sensitive than rEEC in the subgroups of patients with bulbar onset and in patients with disease duration of 12 months or more, the former finding confirming previous observations (de Carvalho and Swash, 2009).

Although 57 patients were upgraded by the AC and only 26 patients were downgraded, this did not result in significantly more patients to include in trials. The main reason for this is that 20 patients were downgraded from "Probable Laboratory-supported" by rEEC to "Possible" by AC, as it has also been found by others (Jang and Bae, 2015; Higashihara et al., 2012). These findings support the suggestion of including a "Probable Laboratory-supported" category in the AC (Geevasinga et al., 2016b).

The diagnostic sensitivities increased, as expected, if patients in the "Possible" category were potentially included in trials as has been suggested (Geevasinga et al., 2016b; Geevasinga et al., 2016a; Berry et al., 2017). This would allow inclusion of only 9 more patients using AC, a non-significant benefit. Our data could not, however, confirm that this increase in sensitivity is gained without a decrease in specificity (Table 2).

4.5. Possible explanations for inter-rater variation

One possible explanation for the large inter-rater variation on both the rEEC and the AC is their complexity. The complexity of the rEEC is inherent in the AC, as the AC were constructed as an electrophysiological amendment to the rEEC. Some of the factors leading to the complexity of the rEEC criteria are summarised in Table 3. The rEEC definition delimits four body regions; bulbar, cervical, thoracic, and lumbosacral, with the cervical and lumbosacral regions including right and left sided limbs, and with different requirements for the number of abnormal muscles in one region, i.e. only one in the bulbar region. Another requirement, the importance of UMN signs rostral to LMN signs for the "Probable" category, adds to the complexity. This requirement was introduced to exclude patients with spinal lesions, but this may now be obsolete with today's availability of MRI for exclusion of spinal pathology; the rEEC require imaging only to exclude other diseases.

Table 3
Some factors that may add to the complexity of revised El Escorial criteria.

- Definition of four body regions
- Different requirements on number of abnormal muscles in the body regions
- UMN signs rostral to LMN signs for some categories
- UMN and LMN signs required in the same region or in different regions
- Concomitant acute and chronic EMG changes required in a muscle
- Interpretation of EMG findings as acute or chronic by the use of different kinds of examination techniques

Both for the rEEC and the AC, it is not described consistently whether UMN and LMN signs are required in the same region or in different regions. The requirement for the presence of concomitant signs of acute and chronic denervation on EMG also adds to the complexity, as this requires detailed knowledge on the pathophysiological interpretation of many different electrophysiological findings. Many neurophysiologists may not be familiar with or accept others' examination techniques (Johnsen et al., 1995). In particular, finding sparse fasciculation potentials in a muscle can be time consuming and requires experience. It has been found that muscle ultrasound findings increase the sensitivity of AC (Misawa et al., 2011; Grimm et al., 2015).

Another factor that may contribute to inter-rater variation is the criteria's format. The rEEC and the AC were both originally published in very detailed formats (Brooks et al., 2000; de Carvalho et al., 2008), and there is no "official flow-chart" that can be used for a quick overview in practice. When using the AC, it is required that both documents of rEEC and AC are considered together. To obtain an overview, readers may be tempted to rely on unofficial, not peer-reviewed flow-charts with an inherent risk of errors.

4.6. Upper motor neuron signs

Definitions on UMN signs, applicable for both the rEEC and the AC, are provided in the rEEC (Brooks et al., 2000). However, detection of UMN signs rely on a clinical examination, which is influenced by clinical skills. Objective detection of UMN signs in atrophic limbs poses difficulties (Swash, 2012). Such differences in the clinical detection of UMN signs constitute a possible source of inter-rater variation in diagnostic criteria for ALS. This is probably relevant in practice, but not in the present study, because presence or absence of UMN signs were presented for all raters in the same standardised data sheets.

Concerning improvements in diagnostic sensitivity and specificity, the development of a reliable objective method for detection of UMN signs is probably the most critical item. Threshold tracking transcranial magnetic stimulation is a promising technique to detect UMN involvement. It has been shown to differentiate between ALS and non-ALS disorders early in the disease course (Menon et al., 2015), however, the findings still have to be replicated in larger studies by other groups. Conventional magnetic resonance imaging shows only sparse changes in ALS, which cannot serve as a biomarker for ALS. Advanced neuroimaging techniques such as diffusion tensor imaging and magnetic resonance spectroscopy can show structural and pathophysiological changes in ALS, but these techniques have still not been established as clinical tools for the detection of UMN involvement (Foerster et al., 2013).

4.7. Study limitations

It is likely that higher Kappa values for agreement might have been found if more cases without ALS were included in our study, as the Kappa coefficient is dependent on the incidence (Maclure and Willett, 1987).

The raters of this study interpreted other physician's studies solely from EMG and NCS parameters as written in the standardised data sheets, i.e. no curves were provided. There were no attempts to make a consensus on definitions. For example, the examining physician could state whether fasciculation potentials were stable or unstable, and the rater had to accept this independently on the definition of stability of fasciculation potentials used. The use of the terms "borderline increased/decreased" without a clear definition can have led to individual interpretations, which may have lowered inter-rater agreement.

The participants submitted different numbers of cases, which may have biased the results towards data from the few centres providing most cases. Although all raters had some experience with clinical diagnosing of ALS patients, the level of experience was not controlled in this study as it might have been in a clinical trial. Inter-rater variation might be reduced by pre-study training and quality control at individual laboratories, as recommended in a clinical trial setting. Moreover, the rEEC and AC criteria require a formalised approach to the diagnostic investigation, something that is not necessarily standard practice among neurophysiologists. However, the aim of our study was to show how practical and useful the rEEC and the AC are when they are used in the real world. It should be emphasized, that these criteria were derived as research criteria, not as criteria for diagnosis in ordinary clinical practice.

5. Conclusions

The new concepts of the AC increase diagnostic sensitivity, but this gain is reduced by the omission of the "Probable Laboratory-supported" category. The present study supports previous suggestions of including the "Probable Laboratory-supported" category in an updated version of the AC.

Our results show a marked inter-rater variation in the use of both rEEC and AC most pronounced for classification in subcategories. This variation is most likely due to a large complexity of the rEEC, which is inherent in the AC.

Training might reduce inter-rater variation. However, our results highlight a need for developing more simple and reproducible criteria for use in research and perhaps also in clinical practice.

Conflict of interest statement

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

Appendix A. Supplementary material

An example of a case with some disagreement as presented in the standardised data sheets. The diagnoses according to the revised El Escorial Criteria (rEEC) and to the Awaji Criteria (AW) entered in the sheet are given by physician who did the studies. These diagnoses were removed from the sheets before they were sent to eight physicians who were prompted to give their diagnoses. The eight raters' diagnoses according to the rEEC were: 1 Definite, 1 Probable, 4 Probable Laboratory-supported, and 2 Possible. Diagnoses according to the AC were: 2 Definite, 2 Probable, and 4 Possible. Supplementary data to this article can be found online at https://doi.org/10.1016/j.clinph.2018.11.021.

References

Berry JD, Paganoni S, Atassi N, Macklin EA, Goyal N, Rivner M, Simpson E, Appel S, Grasso DL, Mejia NI, Mateen F, Gill A, Vieira F, Tassinari V, Perrin S. Phase IIa trial of fingolimod for amyotrophic lateral sclerosis demonstrates acceptable acute safety and tolerability. Muscle Nerve 2017;56:1077–84.

Boekestein WA, Kleine BU, Hageman G, Schelhaas HJ, Zwarts MJ. Sensitivity and specificity of the 'Awaji' electrodiagnostic criteria for amyotrophic lateral sclerosis: retrospective comparison of the Awaji and revised El Escorial criteria for ALS. Amyotroph Lateral Scler 2010;11:497–501.

Brooks BR, Miller RG, Swash M. Munsat TL, World Federation of Neurology Research Group on Motor Neuron Diseases. El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. Amyotroph Lateral Scler Other Motor Neuron Disord 2000;1:293–9.

Brooks BR, Antel J, Bradley W, Cardy P, Carpenter S, Chou S, Conradi S, Daube J, Denys EH, Festoff B, Hirano A, Hormigo A, Karpati G, Kasarskis E, Kuther G, Larumbe R, Leigh N, Martinez-Lage J, Meininger V. El Escorial World Federation of Neurology criteria for the diagnosis of amyotrophic lateral sclerosis. J Neurol Sci 1994;124:96–107.

Costa J, Swash M, de Carvalho M. Awaji criteria for the diagnosis of amyotrophic lateral sclerosis:a systematic review. Arch Neurol 2012;69:1410–6.

de Carvalho M, Dengler R, Eisen A, England JD, Kaji R, Kimura J, Mills K, Mitsumoto H, Nodera H, Shefner J, Swash M. Electrodiagnostic criteria for diagnosis of ALS. Clin Neurophysiol 2008;119:497–503.

de Carvalho M, Swash M. Awaji diagnostic algorithm increases sensitivity of El Escorial criteria for ALS diagnosis. Amyotroph Lateral Scler 2009;10:53–7.

Fleiss JL. Measuring Nominal Scale Agreement Among Many Raters. Psychol Bull 1971;76:378–82.

Foerster BR, Welsh RC, Feldman EL. 25 Years of neuroimaging in amyotrophic lateral sclerosis. Nat Rev Neurol 2013;9:513–24.

Forbes RB, Colville S, Swingler RJ. Are the El escorial and revised El escorial criteria for ALS reproducible? a study of inter-observer agreement. Amyotroph Lateral Scler Other Motor Neuron Disord 2001;2:135–8.

Geevasinga N, Menon P, Scherman DB, Simon N, Yiannikas C, Henderson RD, Kiernan MC, Vucic S. Diagnostic criteria in amyotrophic lateral sclerosis: a multicenter prospective study. Neurology 2016a;87:684–90.

Geevasinga N, Loy CT, Menon P, de Carvalho M, Swash M, Schrooten M, Van Damme P, Gawel M, Sonoo M, Higashihara M, Yi Noto, Kuwabara S, Kiernan MC, Macaskill P, Vucic S. Awaji criteria improves the diagnostic sensitivity in amyotrophic lateral sclerosis: a systematic review using individual patient data. Clin Neurophysiol 2016b;127:2684–91.

Grimm A, Prell T, Decard BF, Schumacher U, Witte OW, Axer H, Grosskreutz J. Muscle ultrasonography as an additional diagnostic tool for the diagnosis of amyotrophic lateral sclerosis. Clin Neurophysiol 2015;126:820–7.

Higashihara M, Sonoo M, Imafuku I, Fukutake T, Kamakura K, Inoue K, Hatanaka Y, Shimizu T, Tsuji S, Ugawa Y. Fasciculation potentials in amyotrophic lateral sclerosis and the diagnostic yield of the awaji algorithm. Muscle Nerve 2012;45:175–82.

Jang JS, Bae JS. Awaji criteria are not always superior to the previous criteria: a meta-analysis. Muscle Nerve 2015;51:822–9.

Johnsen B, Fuglsang-Frederiksen A, Vingtoft S, Fawcett P, Liguori R, Nix W, Otte G, Schofield I, Veloso M, Vila A. Inter- and intraobserver variation in the interpretation of electromyographic tests. Electroencephalogr Clin Neurophysiol 1995;97:432–43.

Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126:161–9.

Menon P, Geevasinga N, Yiannikas C, Howells J, Kiernan MC, Vucic S. Sensitivity and specificity of threshold tracking transcranial magnetic stimulation for diagnosis of amyotrophic lateral sclerosis: a prospective study. Lancet Neurol. 2015;14:478–84.

Misawa S, Noto Y, Shibuya K, Isose S, Sekiguchi Y, Nasu S, Kuwabara S. Ultrasonographic detection of fasciculations markedly increases diagnostic sensitivity of ALS. Neurology 2011;77:1532–7.

Okita T, Nodera H, Shibuta Y, Nodera A, Asanuma K, Shimatani Y, Sato K, Izumi Y, Kaji R. Can Awaji ALS criteria provide earlier diagnosis than the revised El Escorial criteria? J Neurol Sci 2011;302:29–32.

Ross MA, Miller RG, Berchert L, Parry G, Barohn RJ, Armon C, Bryan WW, Petajan J, Stromatt S, Goodpasture J, McGuire D. Toward earlier diagnosis of amyotrophic lateral sclerosis: revised criteria. rhCNTF ALS study group. Neurology 1998;50:768–72.

Swash M. Why are upper motor neuron signs difficult to elicit in amyotrophic lateral sclerosis? J Neurol Neurosurg Psychiatry 2012;83:659–62.

Traynor BJ, Codd MB, Corr B, Forde C, Frost E, Hardiman OM. Clinical features of amyotrophic lateral sclerosis according to the El Escorial and Airlie House diagnostic criteria: a population-based study. Arch. Neurol. 2000;57:1171–6.

Turner MR, Kiernan MC, Leigh PN, Talbot K. Biomarkers in amyotrophic lateral sclerosis. Lancet Neurol 2009;8:94–109.

Vingtoft S, Johnsen B, Fuglsang-Frederiksen A, Veloso M, Barahona P, Vila A, Fawcett P, Scofield I, Ladegaard J, Otte G. ESTEEM: a European telematic project for quality assurance within clinical neurophysiology. Medinfo 1995;8:1047–51.

Zoccolella S, Beghi E, Palagano G, Fraddosio A, Samarelli V, Lamberti P, Lepore V, Serlenga L, Logroscino G. Predictors of delay in the diagnosis and clinical trial entry of amyotrophic lateral sclerosis patients: A population-based study. J Neurol Sci 2006;250:45–9.

Zou G, Donner A. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. Biometrics 2004;60:807–11.