OPEN ACCESS Freely available online

Tremor and Other Hyperkinetic Movements

Brief Reports

Inter-Rater Agreement in the Clinical Diagnosis of Essential Tremor: Data from the NEDICES-2 Pilot Study

Fernando Sierra-Hidalgo[1,2*], Juan Pablo Romero[2], Felix Bermejo-Pareja[1,3,4], Alvaro Sánchez-Ferro[2], Jesús Hernández-Gallego[1,4], Ignacio J. Posada[1,4], Julián Benito-León[1] & Elan D. Louis[5,6,7,8], for the NEDICES-2 Neurological Team[†]

[1] Neurology Department, University Hospital "12 de Octubre", Madrid, Spain, [2] Research Institute, University Hospital "12 de Octubre" (i+12), Madrid, Spain, [3] Centro de Investigación Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED), Madrid, Spain, [4] Department of Medicine, Faculty of Medicine, Universidad Complutense, Madrid, Spain, [5] Gertrude H Sergievsky Center, Columbia University, New York, New York, United States of America, [6] Department of Epidemiology, Columbia University, New York, New York, United States of America, [7] Department of Neurology, Columbia University, New York, New York, United States of America, [8] Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, New York, United States of America, [†] The NEDICES-2 Neurological Team: Julián Benito-León, Félix Bermejo-Pareja, Esther Cubo, Jesús Hernández-Gallego, Jaime Herreros, Andrés Labiano-Fontcuberta, Elan D. Louis, José Antonio Molina, Ignacio Javier Posada, Juan Pablo Romero-Muñoz, Álvaro Sánchez-Ferro, Fernando Sierra-Hidalgo, Alberto Villarejo-Galande.

Abstract

Background: Our aim was to assess the diagnostic agreement among the neurologists in the Neurological Disorders in Central Spain 2 (NEDICES-2) study; these neurologists were assigning diagnoses of essential tremor (ET) vs. no ET.

Methods: Clinical histories and standardized video-taped neurological examinations of 26 individuals (11 ET, seven Parkinson's disease, three diagnostically unclear, four normal, one with a tremor disorder other than ET) were provided to seven consultant neurologists, six neurology residents, and five neurology research fellows (18 neurologists total). For each of the 26 individuals, neurologists were asked to assign a diagnosis of "ET" or "no ET" using diagnostic criteria proposed by the Movement Disorders Society (MDS). Inter-rater agreement was assessed both with percent concordance and non-weighted κ statistics.

Results: Overall κ was 0.61 (substantial agreement), with no differences between consultant neurologists (κ=0.60), neurology residents (κ=0.61), and neurology research fellows (κ=0.66) in subgroup analyses. Subanalyses of agreement only among those 15 subjects with a previous diagnosis of ET (11 patients) and those with a previous diagnosis of being normal (four individuals) showed an overall κ of 0.51 (moderate agreement).

Discussion: In a population-based epidemiological study, substantial agreement was demonstrated for the diagnosis of ET among neurologists of different levels of expertise. However, agreement was lower than that previously reported using the Washington Heights–Inwood Genetic Study of Essential Tremor criteria, and a head-to-head comparison is needed to assess which is the tool of choice in epidemiological research in ET.

Keywords: Tremor, essential tremor, clinical diagnosis, inter-rater agreement, reliability

Citation: Sierra-Hidalgo F, Romero JP, Bermejo-Pareja F, et al. Inter-rater agreement in the clinical diagnosis of essential tremor: Data from the NEDICES-2 Pilot Study. Tremor Other Hyperkinet Mov 2014; 4. doi: 10.7916/D8JD4TQ0

*To whom correspondence should be addressed. E-mail: fsierra.hdoc@salud.madrid.org

Editor: Ruth Walker, James J. Peters Veterans Affairs Medical Center, United States of America

Received: June 11, 2013 **Accepted:** December 29, 2013 **Published:** February 4, 2014

Funding: Fondo de Investigaciones Sanitarias (FIS) del Instituto de Salud Carlos III (ISCIII): PI10/02937, PI11/1508, PI12/01602. Dr. Fernando Sierra-Hidalgo and Dr. Álvaro Sánchez-Ferro receive financial support from the ISCIII through a "Rio-Hortega" contract. Drs. Félix Bermejo-Pareja and Julián Benito-León are supported by NIH R01 NS039422 from the National Institutes of Health, Bethesda, MD, USA. Dr. Elan D. Louis is supported by R01 NS042859 and R01 NS039422 from the National Institutes of Health, Bethesda, MD, USA.

Financial Disclosures: None.

Conflict of Interest: The authors report no conflict of interest.

Introduction

The ideal gold standard for the diagnosis of a disease is an easily identifiable pathological finding or, in the absence of this, a disease-specific biological marker.[1] The absence of biomarkers or diagnostic pathological findings for many neurological disorders adds uncertainty to their diagnosis.[2]

For this reason, the diagnosis of these neurological disorders relies on expert clinical assessment, using previously established diagnostic criteria.[3,4] Thus, it is critically important to determine the diagnostic agreement among experts. Inter-rater agreement in the diagnosis of essential tremor (ET) has been previously assessed by Louis et al.[5] in a study of 226 subjects, which demonstrated a diagnostic concordance of 80% and a weighted κ statistic of 0.84 between two neurologists specializing in movement disorders who used the Washington Heights–Inwood Genetic Study of Essential Tremor (WHIGET) protocol and clinical criteria.

The Neurological Disorders in Central Spain 2 (NEDICES-2) is a population-based, closed cohort study that will assess over 10,000 subjects from several populations in central Spain; it will also include a biobank. All participants will be screened and, if necessary, assessed by a neurologist for the presence of several neurological conditions (i.e., Parkinson's disease, ET, mild cognitive impairment, dementia, transient ischemic attacks, stroke, headaches, sleep disorders, and oro-linguo-facial dyskinesia). Our aim here was to perform a reliability study among the participant neurologists with respect to the diagnosis of ET vs. no ET.

Methods

The NEDICES-2 is a population-based epidemiological study, which will include over 10,000 subjects aged 55 years and older from the regions of Madrid, Ávila, Segovia, Burgos, and Salamanca. Face-to-face interviews will include a comprehensive questionnaire on demographics, current medications, medical conditions, and lifestyle habits; biological samples (blood, saliva, urine, and hair) will be obtained at baseline. Presently, the project is in the pilot study phase. The Clinical Research Ethics Committee of the Hospital 12 de Octubre Research Institute has approved the protocol of the NEDICES-2 study and its pilot study.

The work was conducted at the University Hospital 12 de Octubre in Madrid (Spain), which is the tertiary care center coordinating the NEDICES-2 project. Twenty-six patients were selected from the database of the movement disorders clinic of this institution by an independent team of researchers (not involved in this agreement study); the patients had signed informed consent for the research use of their data. The patients were selected in an attempt to cover the wide spectrum of tremor presentations, including severe ET, mild, or moderate ET, unclear tremor diagnosis, and those with no tremor at all (normal). Among the selected patients, there were four individuals with a severe disabling postural and kinetic tremor and a diagnosis of ET ("severe ET" category; cases 1, 8, 9, and 12), seven individuals with previous diagnoses of mild to moderate ET ("mild/moderate ET" category; cases 3, 5, 10, 16, 17, 19, and 23), four individuals with a diagnosis of no tremor or physiological tremor and completely normal neurological examination ("normal" category; cases 4, 13, 21, and 22), seven patients with a diagnosis of Parkinson's disease ("PD" category; cases 7, 11, 14, 20, 24, 25, and 26), one subject with a diagnosis of another tremor different to ET ("other tremor" category; case 18), and three individuals that were considered *a priori* to be diagnostically unclear due to the presence of mild postural and intention tremor along with parkinsonian signs, such as hypomimia and mild bradykinesia ("ET/PD" category; cases 2, 6, and 15). These subjects did not have a definite diagnosis, and the differential included ET and Parkinson's disease.

A questionnaire was mailed to seven consultant neurologists, six neurology residents, and five neurology research fellows (18 neurologists in total) who worked at the Department of Neurology. They were provided a history of the clinical presentation of the 26 subjects and a video-recording of a standardized neurological examination, including assessment of head, trunk, and upper limb tremor at rest, and during sustained arm extension, pouring water, drinking water, and finger-to-nose maneuver. The 18 neurologists were blinded to the diagnosis previously assigned by clinical neurologists with expertise in movement disorders, and independently assessed the information and provided a diagnosis. The possible answers for each subject were "ET" or "no ET", assessed using the diagnostic criteria proposed by the Movement Disorders Society (MDS).[6]

Inter-rater agreement was assessed with concordance (i.e., percentage of 18 neurologists who agreed with the clinic-assigned diagnosis of "ET" or "no ET") and was also analyzed by means of a non-weighted κ statistic for multiple raters with two possible outcomes (Stata 12, Stata Corp, College Station, TX). The κ statistic takes chance agreement into account, whereas concordance does not.[7] κ coefficients were graded as proposed by Landis and Koch:[8] 0–0.2 (slight agreement), 0.21–0.4 (fair agreement), 0.41–0.6 (moderate agreement), 0.61–0.8 (substantial agreement), and 0.81–1.0 (near perfect agreement). Subgroup analyses of inter-rater agreement were also performed depending on the expertise of the 18 neurologists (consultants, research fellows, and residents).

Results

Diagnosis of ET was made by 100% of raters in one subject (case 1 with severe ET), and the diagnosis of "no ET" was made by 100% of raters in six subjects (case 4 [normal], 7, 11, 24, 25, 26 [with PD]) (Table 1). The percentage agreement for diagnostic categories "mild to moderate ET", "severe ET", and "ET/PD" was variable from case to case. Overall, the highest percentage agreement seemed to be achieved in the cases previously rated as "PD", "other tremor", and "normal".

Overall κ was 0.61 (95% CI 0.49–0.64), which is in the range of moderate to substantial agreement (Table 2). Subgroup analyses showed that κ was 0.60 (95% CI 0.57–0.69) among consultant neurologists (moderate to substantial agreement), 0.66 (95% CI 0.52–0.78) among research fellows (moderate to substantial agreement), and 0.61 (95% CI 0.49–0.67) among neurology residents (moderate to substantial agreement). Subanalyses of agreement only among those 15

Table 1. Overall and Subgroup Percent Agreement in the Diagnosis of Essential Tremor

A Priori Diagnosis	Case Numbers	Overall		Consultants		Residents		Research Fellows	
		ET (%)	No ET (%)	ET (%)	No ET (%)	ET (%)	No ET (%)	ET (%)	No ET (%)
Mild/moderate ET	3	88.9	11.1	100.0	0.0	66.7	33.3	100.0	0.0
	5	38.9	61.1	71.4	28.6	0.0	100.0	40.0	60.0
	10	5.6	94.4	100.0	0.0	100.0	0.0	80.0	20.0
	16	72.2	27.8	85.7	14.3	50.0	50.0	80.0	20.0
	17	83.3	16.7	85.7	14.3	83.3	16.7	80.0	20.0
	19	88.9	11.1	100.0	0.0	83.3	16.7	80.0	20.0
	23	88.9	11.1	100.0	0.0	66.7	33.3	100.0	0.0
Severe ET	1	100.0	0.0	100.0	0.0	100.0	0.0	100.0	0.0
	8	55.6	44.4	71.4	28.6	33.3	66.7	60.0	40.0
	9	94.4	5.6	100.0	0.0	100.0	0.0	80.0	20.0
	12	61.1	38.9	71.4	28.6	50.0	50.0	60.0	40.0
ET/PD	2	88.9	11.1	100.0	0.0	66.7	33.3	100.0	0.0
	6	27.8	72.2	57.1	42.9	0.0	100.0	20.0	80.0
	15	11.1	88.9	28.6	71.4	0.0	100.0	0.0	100.0
PD	7	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0
	11	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0
	14	5.6	94.4	14.3	85.7	0.0	100.0	0.0	100.0
	20	5.6	94.4	14.3	85.7	0.0	100.0	0.0	100.0
	24	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0
	25	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0
	26	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0
Other tremor	18	16.7	83.3	28.6	71.4	16.7	83.3	0.0	100.0
Normal	4	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0
	13	5.6	94.4	14.3	85.7	0.0	100.0	0.0	100.0
	21	5.6	94.4	14.3	85.7	0.0	100.0	0.0	100.0
	22	11.1	88.9	28.6	71.4	0.0	100.0	0.0	100.0

Abbreviations: ET, Essential Tremor; PD, Parkinson's Disease.

subjects with a previous diagnosis of ET (11 patients) and those with a previous diagnosis of being normal (4 individuals) showed an overall κ of 0.51 (95% CI 0.44–0.66, Z=24.56, p<0.001), and subgroup analyses showed κ=0.52 among neurologists (95% CI 0.42–0.55, Z=9.24, p<0.001), κ=0.54 among residents (95% CI 0.41–0.65, Z=8.06, p<0.001), and κ=0.48 among research fellows (95% CI 0.39–0.65, Z=5.91, p<0.001); these values were all in the range of moderate agreement.

Discussion

The goal of case identification in epidemiological research is to obtain a standardized diagnosis that is the most accurate possible

Table 2. Overall and Subgroup Diagnostic Agreement

Raters	N	κ	95% CI	Z	p
Overall	18	0.61	0.49–0.64	38.4	<0.001
Neurologists	7	0.60	0.57–0.69	14.12	<0.001
Residents	6	0.61	0.49–0.67	12.11	<0.001
Research fellows	5	0.66	0.52–0.78	10.58	<0.001

Abbreviation: CI, Confidence Interval.

within the constraints of the study design and available resources.[3] The basic tool in neurological diagnosis is expert examination. Even with the expertise of specialists, a definite diagnosis may not be possible in some cases during life.

Misclassification of disease status in epidemiological research dilutes the true association between exposure and disease, when misclassification is random, and may falsely elevate the degree of association between an exposure and disease risk when there is systematic identification bias.[3] For most studies of neurological diseases, routine clinical diagnosis by neurologists, often in conjunction with the use of standardized published diagnostic criteria, is the most practical method for case identification. Standardized diagnostic criteria are imperfect, but can help to ensure that various groups involved in research are in fact studying the same entity. However, routine clinical diagnosis depends on the expertise of the clinician and can be affected by differences in disease presentation and in the attitudes of physicians toward the diagnosis in different cultures.

The current results, among researchers involved in the NEDICES-2 study, indicate that the MDS consensus diagnostic criteria are a reliable set of criteria within the current framework. These results show an overall substantial agreement for the diagnosis of ET, which is similar among neurologists, research fellows, and neurology residents. Subanalyses limited to all severity of ET cases as well controls revealed an overall level of agreement that was lower but still remained in the moderate range. The MDS criteria were selected because of their simplicity and rapid application using data from the medical history and the physical examination of cases. We attempted to minimize the variability in the patient's medical records and examinations by reformatting the data into a standard case record format and a standardized physical examination, and we then required raters to classify cases into diagnostic groups using standardized diagnostic criteria.[9] WHIGET criteria have the benefit of recording a standardized neurological examination and assessing it by means of a previously validated score.[5] However, this scale has the disadvantage of having been validated only among experts in movement disorders. The present study has demonstrated an acceptable rate of agreement among non-specialists using a simpler diagnostic tool. The values of κ are lower than that found in the agreement study by Louis et al.;[5] this could be a function of the different case mix and the different level of expertise of the neurologists in the two studies. It could also reflect the diagnostic tools that were used.

This study has several limitations. Firstly, this was a reliability study, not a validity study. In the absence of biologic markers for ET (i.e., a diagnostic gold standard), the issue of validity becomes a difficult one to address.[5] Reliability becomes the only standard by which one can judge the quality of the observations. Secondly, while we assessed inter-rater agreement, we did not assess test–retest reliability.[10] Third, the use of video-taped examinations may add some concerns. However, Martínez-Martín et al.[11] showed that rating action tremors without the assistance of a teaching video-tape was characterized by only moderate levels of inter-rater agreement. On the other hand, the apparent amplitude of a tremor seen on a video-screen also depends on the distance of the observers from the screen and the size of the images, which is influenced by the amount of zoom used by the cameraman.[12] The accuracy of the video-recording for detecting tremor also depends on the rate of the movement, with information being lost the faster the tremor frequency, and, thus, there is a greater reduction in the apparent amplitudes of high- compared with low-frequency tremors.[12] Fourth, in terms of statistical tests, the κ test can be quite sensitive. Inclusion of only a group of easy to diagnose cases biases the analysis towards a high agreement. An attempt to minimize this effect was made by selecting cases with different severities of tremor, subjects without tremor at all, and subjects with an unclear diagnosis. However, given the variety of diagnosis, the sample size in each category is probably lower than desirable. Therefore, the results of subgroup analyses must be interpreted with caution. Finally, we did not test different plausibility ratings for the diagnosis of ET (i.e., definite, probable, and possible). The distinction between normal and possible ET is still an area of some disagreement, with only moderate agreement between experts.[5]

In summary, we have demonstrated a substantial agreement among neurologists with different levels of expertise involved in a population-based epidemiological study of ET. However, agreement rates were lower than those previously reported using the WHIGET criteria, and a head-to-head comparison is needed to assess which is the tool of choice in epidemiological research in ET. A standardized training session on reliability, with the participation of the researchers to be involved in the clinical assessment of NEDICES-2 participants, would be necessary in order to increase the reliability of the diagnosis of ET if the MDS criteria are to be used.

Acknowledgements

Thanks to J. L. Pons and E. Rocón (Directors of the Neurotremor European Study project) for their advice; to all the General Practice doctors taking part in the pilot study of the NEDICES-2 project; to all the neurologists, research fellows, and neurology residents taking part in the diagnostic agreement study of the NEDICES-2 project, and Rocío Trincado for statistical analyses.

References

1. Louis ED, Pullman SL. Comparison of clinical vs. electrophysiological methods of diagnosing of essential tremor. *Mov Disord* 2001;16:668–673, doi: http://dx.doi.org/10.1002/mds.1144.

2. Rajput AH, Rozdilsky B, Ang L, Rajput A. Clinicopathologic observations in essential tremor: report of six cases. *Neurology* 1991;41:1422–1424, doi: http://dx.doi.org/10.1212/WNL.41.9.1422.

3. Tanner CM, Webster Ross G. Neuroepidemiology: Fundamental considerations. In: Nelson LM, Tanner CM, Van Der Eeden SK, McGuire VM, editors. Neuroepidemiology. From principles to practice. New York: Oxford University Press; 2004. p 1–22.

4. Chouinard S, Louis ED, Fahn S. Agreement among movement disorder specialists on the clinical diagnosis of essential tremor. *Mov Disord* 1997;12:973–976, doi: http://dx.doi.org/10.1002/mds.870120621.

5. Louis ED, Ford B, Bismuth B. Reliability between two observers using a protocol for diagnosing essential tremor. *Mov Disord* 1998;13:287–293, doi: http://dx.doi.org/10.1002/mds.870130215.

6. Deuschl G, Bain P, Brin M. Consensus statement of the Movement Disorder Society on Tremor. Ad Hoc Scientific Committee. *Mov Disord* 1998; 13(Suppl 3):2–23.

7. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;76:378–382, doi: http://dx.doi.org/10.1037/h0031619.

8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174, doi: http://dx.doi.org/10.2307/2529310.

9. Baldereschi M, Amato MP, Nencini P, et al. Cross-national interrater agreement on the clinical diagnostic criteria for dementia. WHO-PRA Age-Associated Dementia Working Group, WHO-Program for Research on Aging, Health of Elderly Program. *Neurology* 1994;44:239–242, doi: http://dx.doi.org/10.1212/WNL.44.2.239.

10. Louis ED, Barnes LF, Wendt KJ, et al. Validity and test-retest reliability of a disability questionnaire for essential tremor. *Mov Disord* 2000;15:516–523, doi: http://dx.doi.org/10.1002/1531-8257(200005)15:3<516::AID-MDS1015>3.0.CO;2-J.

11. Martínez-Martín P, Gil-Nagel A, Gracia LM, Gómez JB, Martínez-Sarriés J, Bermejo F. Unified Parkinson's Disease Rating Scale characteristics and structure. The Cooperative Multicentric Group. *Mov Disord* 1994;9:76–83, doi: http://dx.doi.org/10.1002/mds.870090112.

12. Bain PG, Findley LJ, Atchison P, et al. Assessing tremor severity. *J Neurol Neurosurg Psychiatry* 1993;56:868–873, doi: http://dx.doi.org/10.1136/jnnp.56.8.868.