Interrater reliability of EEG-video monitoring

S.R. Benbadis, MD
W.C. LaFrance, Jr., MD,
MPH
G.D. Papandonatos, PhD
K. Korabathina, MD
K. Lin, MD
H.C. Kraemer, PhD
For the NES Treatment
Workshop*

Address correspondence and
reprint requests to Dr. Selim R.
Benbadis, 2 Tampa General
Circle, 7th Floor, Tampa, FL
33606
sbenbadi@health.usf.edu

ABSTRACT

Objective: The diagnosis of psychogenic nonepileptic seizures (PNES) can be challenging. In the absence of a gold standard to verify the reliability of the diagnosis by EEG-video, we sought to assess the interrater reliability of the diagnosis using EEG-video recordings.

Methods: Patient samples consisted of 22 unselected consecutive patients who underwent EEG-video monitoring and had at least an episode recorded. Other test results and histories were not provided because the goal was to assess the reliability of the EEG-video. Data were sent to 22 reviewers, who were board-certified neurologists and practicing epileptologists at epilepsy centers. Choices were 1) PNES, 2) epilepsy, and 3) nonepileptic but not psychogenic ("physiologic") events. Interrater agreement was measured using a κ coefficient for each diagnostic category. We used generalized κ coefficients, which measure the overall level of between-method agreement beyond that which can be ascribed to chance. We also report category-specific κ values.

Results: For the diagnosis of PNES, there was moderate agreement ($\kappa = 0.57$, 95% confidence interval [CI] 0.39–0.76). For the diagnosis of epilepsy, there was substantial agreement ($\kappa = 0.69$, 95% CI 0.51–0.86). For physiologic nonepileptic episodes, the agreement was low ($\kappa = 0.09$, 95% CI 0.02–0.27). The overall κ statistic across all 3 diagnostic categories was moderate at 0.56 (95% CI 0.41–0.73).

Conclusions: Interrater reliability for the diagnosis of psychogenic nonepileptic seizures by EEG-video monitoring was only moderate. Although this may be related to limitations of the study (diagnosis based on EEG-video alone, artificial nature of the forced choice paradigm, single episode), it highlights the difficulties and subjective components inherent to this diagnosis.

Neurology® 2009;73:843–846

GLOSSARY

ABCN = American Board of Clinical Neurophysiology; **ABPN** = American Board of Psychiatry and Neurology; **CI** = confidence interval; **IRR** = interrater reliability; **PNES** = psychogenic nonepileptic seizures.

Psychogenic nonepileptic seizures (PNES) are episodes that resemble epileptic seizures but have a psychological origin.¹ Many transient neurologic symptoms can be misdiagnosed as epilepsy, including syncope, movement disorders, and parasomnias, but PNES are by far the most common at epilepsy centers. The gold standard for diagnosis of PNES is generally considered to be EEG-video monitoring, but its accuracy is unknown because there is no confirmatory test, such as pathology, and intracranial electrodes carry significant risks. In the absence of a definitive confirmatory gold standard, interrater agreement may be the best measure of diagnostic reliability. Based on benchmarks from the National Institute of Neurological Disorders and Stroke/National Institute of Mental Health/American Epilepsy Society–sponsored nonepileptic seizures treatment workshop,² this study sought to evaluate interrater reliability (IRR) for the diagnosis of seizures based on EEG-video monitoring.

*See the appendix for information about the NES Treatment Workshop.

From the University of South Florida and Tampa General Hospital (S.R.B., K.K., K.L.), Tampa, FL; Rhode Island Hospital, Brown Medical School (W.C.L.), Providence, RI; Brown University (G.D.P.), Providence, RI; and Stanford University (H.C.K.), Palo Alto, CA.

The NES Treatment Workshop was sponsored by the National Institute of Neurological Disorders and Stroke, the National Institute of Mental Health, and the American Epilepsy Society.

Disclosure: Author disclosures are provided at the end of the article.

Preliminary data were presented as an abstract at the 2007 American Epilepsy Society, Philadelphia, PA.

METHODS Standard protocol approvals, registrations, and patient consent. The study was approved by the institutional review board at the University of South Florida and Tampa General Hospital. Written informed consent for education and research was obtained from all patients (or guardians of patients) participating in the study.

Patient samples were collected at 1 center (University of South Florida and Tampa General Hospital) and consisted of 22 unselected consecutive patients who underwent noninvasive EEG-video monitoring and had at least 1 episode recorded. Data were collected in a standard fashion that included interictal samples and all recorded episodes. The standard 10–20 electrode system was used, including the T1 and T2 electrodes (total 23 electrodes). Recordings were acquired as a double banana but were readily reformattable to be viewed in different montages at the reviewer's preference. Each patient vignette included samples of interictal EEG (unmarked) and a single recorded episode. EKG was recorded. To approximate the clinical scenario, the sample provided for the reviewers was the same as what is typically saved for patients undergoing EEG-video.

Data were recorded on XLTEK (Natus Medical, San Carlos, CA, and Ontario, Canada), stored on a DVD, and sent to 22 reviewers. Each rater reviewed all 22 vignettes. Age, sex, and the video EEG-video sample were the only information provided for the reviewer. Results of other tests (e.g., imaging) and extensive histories were not provided, because the goal was to assess the reliability of the EEG-video data for interpretation.

Reviewers were board-certified neurologists and practicing epileptologists at epilepsy centers. The 22 readers comprised 19 from across the United States and 3 from Europe. All of the US epileptologists were certified by the American Board of Psychiatry and Neurology (ABPN), and 18 had either ABPN neurophysiology added qualification or American Board of Clinical Neurophysiology (ABCN) certification. The 22 epileptologists had a mean of 13 years' postfellowship experience (range 3–33 years, SD 7.3 years).

Choices were 1) PNES, 2) epilepsy, and 3) nonepileptic but not psychogenic ("physiologic" or "organic") events. Interrater agreement was measured using a κ coefficient for each diagnostic category. We used generalized κ coefficients,^{3,4} which measure the overall level of between-method agreement beyond that which can be ascribed to chance. We also report category-specific κ values.

Kappa coefficients have their range constrained by differences in prevalence between the dichotomous measures under investigation, and caution should be exercised in their interpretation when the associated sign test is significant.⁵ In the absence of prevalence differences, standard cutoffs for measuring agreement have been established by Landis and Koch,⁶ which rate them as follows: 0.80–1.00, almost perfect; 0.60–0.80, substantial; 0.40–0.60, moderate; 0.20–0.40, fair; 0.00–0.20, slight; and <0.00, poor.

Confidence interval (CI) estimation was based on a nonparametric bootstrap procedure.⁷ Samples of 22 physicians were drawn with replacement 10,000 times from our data set, followed by random draws of 22 patient ratings provided by these particular physicians, also sampled with replacement. The resulting CIs reflect both physician-level and patient-level variability and are thus appropriate for inference on a wider population of physicians comparable to those recruited in our study, rather than being restricted to this particular group of physicians.

RESULTS Diagnoses by reviewers are shown in the table. All 22 reviewers scored each of the 22 EEG-

Table	Patient categorization by epileptologists' diagnostic choice		
	E	P	O
Vignette	22	0	0
2	22	0	0
3	22	0	0
4	22	0	0
5	21	1	0
6	20	2	0
7	20	1	1
8	19	1	2
9	18	3	1
10	17	4	1
11	17	4	1
12	15	5	2
13	10	3	9
14	3	18	1
15	3	14	5
16	2	15	5
17	0	20	2
18	0	19	3
19	0	16	6
20	0	19	3
21	0	21	1
22	0	21	1

For 22 vignettes of EEG-video, rated by 22 epileptologists. Data are presented as patient frequency (count).

E = epilepsy; P = psychogenic nonepileptic seizures; O = physiologic nonepileptic event.

video vignettes. Averaging across raters, the percentages in each of the diagnostic categories were as follows: epileptic, 52%; PNES, 39%; and physiologic, 9%. For the diagnosis of PNES, there was moderate agreement ($\kappa = 0.57$, 95% CI 0.39–0.76). For the diagnosis of epilepsy, there was substantial agreement ($\kappa = 0.69$, 95% CI 0.51–0.86). For physiologic nonepileptic episodes, the agreement was low ($\kappa = 0.09$, 95% CI 0.02–0.27). The overall κ statistic was moderate at 0.56 (95% CI 0.41–0.73).

DISCUSSION Our study demonstrated moderate IRR for identifying PNES by EEG-video alone. This finding may seem a little lower than expected, but we propose a few explanations. First, the diagnosis here was, intentionally but artificially, based solely on EEG-video recordings. This of course does not reflect clinical reality, where the actual diagnosis of PNES is made by a combination of patient history (neurologic and psychiatric), examination, and EEG-video monitoring. This process amounts to "knowing the patient." This clinical "knowledge" may be

subjective and difficult to measure, but our findings would suggest that obtaining the complete picture of the patient may be an important part of this diagnosis. Conversely, as found in another study,⁸ diagnosis of seizures by history alone may not be sufficient. Epileptologists' sensitivity for seizure identification was 96% (95% CI 92%–98%), but specificity was only 50% (95% CI 22%–79%). According to the authors, epileptologists rarely miss epileptic seizures (high sensitivity) but more often overcall nonepileptic events as epileptic seizures (low specificity). A follow-up study reflecting current practice, incorporating the combination of these diagnostic elements, would likely increase the κ significantly. To our knowledge, no other study has analyzed IRR of PNES or ES by EEG-video, alone or with the addition of patient history. The only remotely close study was one on routine EEG based on a very brief segment, and variation was “considerable.”⁹ The κ coefficients for IRR can vary dramatically across different fields. As a reference point, one study revealed an IRR of 0.83 between epilepsy centers on whether to perform epilepsy surgery,¹⁰ and the IRR between sleep centers for scoring 5 different sleep stages was 0.68.¹¹ It is well known that the range of κ values is constrained by the margins. Given that only 9% of the ratings fell in the physiologic category in our study, it is no surprise that κ was so low for this category.

Second, there was only 1 episode for each patient, whereas in clinical practice multiple episodes are usually recorded if available and can be important for informing the diagnosis. Third, the “forced” choice of 3 options may also be viewed as artificial, because in clinical practice clinicians occasionally remain diagnostically uncommitted. Although we considered having an “uncertain/unclear” category, we were dissuaded from including this choice for statistical reasons, because this category would have “absorbed” too many patients and made the data uninterpretable. Fourth, it could be argued that the category of physiologic nonepileptic was responsible for most of the disagreement, and the agreement slightly improved (0.64) in a post hoc analysis when excluding the physiologic category. However, the calculated coefficient based on diagnostic category removal is not methodologically valid because we do not know how raters would have behaved if their options had been forced only to a binary epilepsy vs PNES diagnosis. Fifth, a closer look at the data (table) reveals that in 12 of the patients, there was agreement among 19 or more of the 22 reviewers, and in 17 of the patients, there was agreement among 17 or more of the reviewers. This would suggest that the diagnosis is not difficult in most patients, but that there are a few

difficult ones that account for an only moderate overall agreement here.

The study was expected to produce CI lengths slightly in excess of 0.30 for category-specific κ values. This compares well with realized values of 0.37 for the epileptic category and 0.35 for PNES. Because patients with PNES are common at epilepsy centers, additional precision in the estimates would have been gained by increasing the number of patients. To generate a representative sample from the population of interest and to reflect actual practice, we used consecutive unselected patients rather than equal proportions of the diagnostic categories.

Our findings suggest that the diagnosis of PNES continues to represent a challenge, and perhaps also indicate that the “art” of medicine or a subjective component to the diagnosis of seizures is part of neurologic practice. The findings underscore the need for training in identification and distinction of brain-behavior disorders. Last, additional research is needed to delineate diagnostic accuracy and reliability in a full and more realistic clinical setting, i.e., using EEG-video in the context of other data.

AUTHOR CONTRIBUTIONS

Statistical analyses were performed by George D. Papandonatos, PhD, and Helena C. Kraemer, PhD.

DISCLOSURE

Dr. Benbadis serves on scientific advisory boards and speakers' bureaus for Abbott, Cyberonics, GlaxoSmithKline, OrthoMcNeil, Pfizer Inc., Sleepmed-DigiTrace, and UCB Pharma; serves on the editorial board of *Epilepsy and Behavior*, *European Neurology*, *Expert Review of Neurotherapeutics*, and *Epileptic Disorders*; and is a Chief Editor for *eMedicine*. Dr. LaFrance received speaker honorarium from the Epilepsy Foundation; receives research support as Principal Investigator from the NIH [NINDS 1K23NS45902], the Rhode Island Hospital, the Epilepsy Foundation [122982], and the Siravo Foundation; and has acted as consultant for Disability Services. Dr. Papandonatos, Dr. Korabathina, and Dr. Lin report no disclosures. Dr. Kraemer serves/has served on scientific advisory boards for the NIMH Advisory Council and the DSM V Task Force; serves/has served as an Associate Editor of *Statistics in Medicine*, *Psychological Methods*, the *International Journal of Eating Disorders*, the *Journal of Child and Adolescent Psychopharmacology*, and the *Archives of General Psychiatry*; receives royalties from publishing *How Many Subjects?* (1988), *Evaluating Medical Tests* (1992), both Sage Publications, and *To Your Health*, (2005), Oxford University Press; and has received consulting fees in the last year from the NIMH Advisory Council, Stanford University, University of California at San Diego, University of Pittsburgh, and Wesleyan University.

APPENDIX

NES Treatment Workshop committee: W. Curt LaFrance, Jr. (chair), Kenneth Alper, Debra Babcock, John J. Barry, Selim Benbadis, Rochelle Caplan, John Gates, Margaret Jacobs, Andres M. Kanner, Roy Martin, Lynn Rundhaugen, Randy Stewart, and Christina Vert.

NES Treatment Workshop participants: Donna Joy Andrews, Joan Austin, Richard Brown, Brenda Burch, John Campo, Paul Desan, Michael First, Peter Gilbert, Laura Goldstein, Jonathan Halford, Mark Hallett, Cynthia Harden, Gabor Keitner, Helena Kraemer, Roberto Lewis-Fernandez, Gregory Mahr, Claudia Moy, Greer Murphy, Sigita Plioplys, Mark Rusch, Chris Sackellares, Steve Schachter, Patricia Shafer, Daphne Simeon, David Spiegel, Linda Street, Michael Trimble, Valerie Voon, Elaine Wyllie, and Charles Zaroff. Orrin Devinsky, Frank Gilliam,

Dalma Kalogjera-Sackellares, John Mellers, and Markus Reuber contributed significantly before the workshop but were unable to attend.

The following contributors served as EEG-video reviewers (listed alphabetically): Ann M. Bergin, MB, MRCP, Harvard University, Children's Hospital, Boston, MA; Andrew S. Blum, MD, PhD, Rhode Island Hospital, Brown University, Providence, RI; Edward B. Bromfield, MD, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; Bradley J. Davis, MD, Delta Waves Sleep Disorders Center, Colorado Springs, CO; Edward Donnelly, MD, Brown University, Providence, RI; Barbara Dworetzky, MD, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; Stephan Eisenschenk, MD, University of Florida, Gainesville, FL; Eric B. Geller, MD, Institute for Neurology and Neurosurgery at Saint Barnabas, Livingston, NJ; Jonathan J. Halford, MD, Medical University of South Carolina, Charleston, SC; Jay H. Harvey, DO, Texas Epilepsy Group, Dallas, TX; Pongkiat Kankirawatana, MD, University of Alabama Birmingham, AL; Fumisuke Matsuo, MD, University of Utah, Salt Lake City, UT; J. Layne Moore, MD, Ohio State University, Columbus, OH; William J. Nowack, MD, University of Kansas, Kansas City, KS; Markus Reuber, MD, PhD, FRCP, University of Sheffield, UK; Joseph Sirven, MD, Mayo Clinic, Scottsdale, AZ; Christopher T. Skidmore, MD, Thomas Jefferson University, Philadelphia, PA; Brien Smith, MD, Henry Ford Hospital, Detroit, MI; Dragoslav Sokic, MD, PhD, Institute of Neurology, Clinical Centre of Serbia, Belgrade, Serbia; Erik K. St. Louis, MD, Mayo Clinic, Rochester, MN; Willam O. Tatum IV, DO, University of South Florida, Tampa, FL; and Nikola Vojvodic, MD, Institute of Neurology, Clinical Centre of Serbia, Belgrade, Serbia.

Received December 20, 2008. Accepted in final form June 22, 2009.

REFERENCES

1. Benbadis SR. Differential diagnosis of epilepsy. *Continuum Lifelong Learning Neurology* 2007;13:48–70.
2. LaFrance WC Jr, Alper K, Babcock D, et al. Nonepileptic seizures treatment workshop summary. *Epilepsy Behav* 2006;8:451–461.
3. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;76:378–382.
4. Fleiss JL, Levin BA, Paik MC. *Statistical Methods for Rates and Proportions*. 3rd ed. Hoboken, NJ: J. Wiley; 2003.
5. Cook RJ. Kappa and its dependence on marginal rates. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. New York: J. Wiley; 1998:2166–2168.
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174.
7. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York: Chapman & Hall; 1993.
8. Deacon C, Wiebe S, Blume WT, McLachlan RS, Young GB, Matijevic S. Seizure identification by clinical description in temporal lobe epilepsy: how accurate are we? *Neurology* 2003;61:1686–1689.
9. Williams GW, Luders HO, Brickner A, Goormastic M, Klass DW. Interobserver variability in EEG interpretation. *Neurology* 1985;35:1714–1719.
10. Haut SR, Berg AT, Shinnar S, et al. Interrater reliability among epilepsy centers: multicenter study of epilepsy surgery. *Epilepsia* 2002;43:1396–1401.
11. Danker-Hopfe H, Kunz D, Gruber G, et al. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J Sleep Res* 2004;13:63–69.

Register Now for 2009 Fall Conference in Las Vegas

Get the latest updates in neurology and practice management at the 2009 Fall Conference in Las Vegas, November 6 through 8, at the Planet Hollywood Resort and Casino.

Earn up to 19.5 AMA PRA Category 1 credits™ in one weekend, to help fulfill your CME requirements as a necessary step towards maintenance of certification. You also will benefit from the intimate atmosphere of a smaller conference size, with ample opportunity for one-on-one networking with some of the top experts in your areas of interest.

Note: The AAN now offers only one regional conference a year; don't miss out! Register by October 5, 2009, to save on registration and hotel costs. Learn more at www.aan.com/fall.