Published in final edited form as:
Neurology. 1993 December ; 43(12): 2526–2530.

Reliability of seizure classification using a semistructured interview

R. Ottman, PhD, **J.H. Lee, MPH**, **W.A. Hauser, MD**, **S. Hong, MPH**, **D. Hesdorffer, PhD**, **N. Schupf, PhD**, **T.A. Pedley, MD**, and **M.L. Scheuer, MD**
From the G.H. Sergievsky Center, and Division of Epidemiology, School of Public Health (Drs. Ottman, Lee, Hauser, Hong, Hesdorffer, and Schupf), and the Department of Neurology (Drs. Hauser, Pedley, and Scheuer), Columbia University, New York; the Epidemiology of Brain Disorders Research Department (Dr. Ottman), New York State Psychiatric Institute, New York; and the Institute for Basic Research in Developmental Disabilities (Dr. Schupf), Staten Island, NY

Abstract

Methods for standardized classification of epileptic seizures are important for both clinical practice and epidemiologic research. In this study, we developed a strategy for standardized classification using a semistructured telephone interview and operational diagnostic criteria. We interviewed 1,957 adults with epilepsy ascertained from voluntary organizations. To confirm and expand the seizure history, we also interviewed a first-degree relative for 67% of subjects and obtained medical records for 59%. Three lay reviewers used all available information to classify seizures. To assess reliability, each reviewer classified a sample of subjects assigned to the others. In addition, an expert physician classified a sample of subjects assigned to two of the reviewers. Agreement was “moderate-substantial” for generalized-onset seizures, both for the comparisons between pairs of lay reviewers and for the neurologist versus lay reviewers. Agreement was “substantial-almost perfect” for partial-onset seizures, both for pairs of lay reviewers and for the neurologist versus lay reviewers. These results suggest that seizures can be reliably classified by lay reviewers, using operational criteria applied to symptoms ascertained in a semistructured telephone interview.

Accurate diagnosis and classification of seizure disorders are essential in both clinical and research contexts, facilitating appropriate treatment selection, validity and precision of research findings, and comparison of findings across studies. Accuracy in classification is especially crucial for genetic research because genetic contributions are likely to differ among different clinical subtypes of epilepsy. Use of standardized methods for data collection and interpretation can improve reliability (consistency) and validity (accuracy) of diagnosis.

Diagnostic inconsistency can result from variability in (1) the criteria used for diagnosis and classification, (2) the information elicited during a clinical encounter, (3) patient response to the same questions at different times, and (4) the interpretation of patient information. The 1981 International Classification of Epileptic Seizures of the International League Against Epilepsy (ILAE) was an important step toward standardized diagnosis because it provided uniform criteria for diagnosis, thus reducing one source of variability.[1] However, methods for reducing the other three sources of variability have seldom been employed.

We developed the Seizure Classification Interview (SCI) for use in the Epilepsy Family Study of Columbia University, our ongoing study of genetic contributions to epilepsy.[2] This semistructured interview was designed to reduce the second problem noted above, ie,

Address correspondence and reprint requests to Dr. Ruth Ottman, G.H. Sergievsky Center, Columbia University, 630 W. 168th Street, New York, NY 10032.

variability in the information elicited during a clinical encounter, and to facilitate diagnosis and classification in a large-scale epidemiologic study in which examination of each patient by an expert neurologist was impractical. We previously reported[3] that seizure classifications based on SCI data agreed well with those of attending physicians with expertise in epilepsy. In the present study, we report on another aspect of reliability, namely agreement between different diagnosticians in interpreting information collected during the interview.

Methods

Data collection

The methods for data collection in the Epilepsy Family Study of Columbia University have been described in detail previously.[2] Briefly, between 1985 and 1988 we ascertained 1,957 subjects with epilepsy who were ≥18 years old (probands) from 10 voluntary organizations for epilepsy. We used semistructured telephone interviews with probands, administered by trained lay interviewers, to collect information on the proband's seizure history, family composition, and family history of seizures and related disorders.

We obtained medical records for 59% of probands. In addition, to confirm and expand the seizure history, for each proband we attempted to administer a similar semistructured telephone interview to a first-degree relative who had witnessed the proband's seizures. We selected the mother for interview whenever available; when she was unavailable, we selected the father or a sibling. Interviews with relatives were completed for 67% of probands (mothers, 48%; fathers or siblings, 19%).

The participation rate for probands was at least 84% and did not differ substantially across agencies. Eighty-seven percent of probands were white, 55% had ≥1 year of college education, and 60% were women. Probands ranged in age from 18 to 82 years and averaged 36 ± 11 (SD) years.

Diagnostic procedures

The seizure history reviewers were three nonphysician research assistants trained to use operational diagnostic criteria developed for application of the 1981 ILAE classification to our data. The reviewers used all available information for diagnosis of each proband (direct interview, interview with relative, and medical record). Training included viewing a videotape of seizures, detailed explanation of the rationale for each interview question, and checking the first 20 to 30 cases reviewed to ensure that the criteria were used correctly. The reviewers were also trained to recognize indications for review by one of the three study neurologists (W.A.H., T.A.P., or M.L.S.) (eg, cases that appeared to have *both* generalized-onset and partial-onset seizures), and such cases were referred to one of them for diagnosis.

All of the 1,957 included probands were confirmed to have epilepsy (ie, ≥2 seizures not associated with acute metabolic or structural insults to the CNS). Seizures were classified according to the 1981 ILAE criteria.[1] The reviewers used three categories to classify patients according to lifetime history of each seizure type: positive, possible, and negative. Each seizure was classified according to the end point of the event. Thus, for example, secondarily generalized seizures preceded by an aura were classified as secondarily generalized only; the category of simple partial seizure was reserved for simple partial seizures that did not evolve either to complex partial or to secondarily generalized. The final distribution of seizure type in probands was 84% partial onset, 12% generalized onset, 1% both partial and generalized onset, and 3% unclassifiable.

To ensure that the three reviewers adhered to the same protocol, we monitored reliability by assigning to each reviewer a 10% random sample of the cases assigned to other reviewers. The

replicate diagnoses were compared at regular "consensus meetings," and any discrepancies were resolved on a case-by-case basis, including review by a study neurologist as required. For analysis of reliability, we used the original diagnoses (prior to modification resulting from the consensus meetings). However, in the tables, we report the number of subjects with each seizure type according to the final diagnosis.

One of the study neurologists also reviewed a sample of cases independently to assess similarity of his interpretation of the data to that of the lay reviewers. Subjects previously diagnosed with generalized-onset seizures were deliberately oversampled to produce a sample containing approximately equal numbers of presumed generalized- and partial-onset seizures.

Statistical analysis

In the present study, we evaluated agreement in seizure classification (1) between pairs of lay reviewers and (2) between the neurologist and two of the lay reviewers. (For the third lay reviewer, the number of cases reviewed by the neurologist was too small to be informative for this comparison.) We used the kappa statistic[4] to assess agreement beyond chance in the diagnosis of each seizure type. As suggested by Landis and Koch,[5] agreement was considered "almost perfect" when $\kappa \geq 0.81$, "substantial" when $0.80 \geq \kappa \geq 0.61$, "moderate" when $0.60 \geq \kappa \geq 0.41$, "fair" when $0.40 \geq \kappa \geq 0.21$, "slight" when $0.20 \geq \kappa \geq 0$, and "poor" when $\kappa < 0$.

We calculated both unweighted and weighted kappas. For unweighted kappas, a case was classified as an agreement only when both reviewers used exactly the same category (positive, possible, or negative). For weighted kappas, credit was given for partial agreement. We used a weight of 1.0 when both reviewers used the same category, 0.75 for "positive-possible," 0.25 for "possible-negative" and 0 for "positive-negative."

As noted above, in the analysis of agreement between the physician and lay reviewers, we oversampled subjects with presumed generalized-onset seizures. Since kappa is sensitive to baseline prevalence, we adjusted each of the two kappas (weighted and unweighted) for the sampling probabilities.[6] The adjusted-weighted kappas account for both sampling probabilities and partial agreements and hence reflect the most accurate assessment of agreement between the neurologist and lay reviewers.

Results

Agreement between pairs of lay reviewers (table 1)

As with the entire study population, a majority of subjects in the sample had partial-onset seizures, and less than one-fourth of subjects had generalized-onset seizures. As a result, the number of patients with generalized nonconvulsive seizures (absence, myoclonic, or atonic seizures) was small.

For all three pairs of lay reviewers, agreement was moderate-almost perfect ($\kappa = 0.46$ to 1.00) for the broad seizure categories, generalized onset and partial onset (table 1). Within partial-onset seizures, agreement for all three pairs was substantial for secondarily generalized seizures ($\kappa = 0.66$ to 0.69) and substantial-almost perfect for complex partial seizures ($\kappa = 0.68$ to 0.87). For simple partial seizures, however, agreement was moderate for two pairs of reviewers and poor for the remaining pair.

Within generalized-onset seizures, agreement was moderate-substantial for generalized tonic-clonic (GTC) ($\kappa = 0.60$ to 0.73), absence ($\kappa = 0.55$ to 0.79), and atonic ($\kappa = 0.59$, 0.66) seizures, and almost perfect ($\kappa > 0.87$) for myoclonic seizures.

In general, weighted kappas were higher than unweighted kappas, since partial credit was given for partial agreement; however, weights did not change the results substantially.

Agreement between the neurologist and lay reviewers (table 2)

Because of the oversampling of subjects with presumed generalized-onset seizures for the comparisons between the neurologist and lay reviewers, 55% (neurologist versus reviewer 1) and 59% (neurologist versus reviewer 2) of subjects in the two comparisons had generalized-onset seizures.

Table 2 shows agreement between the neurologist and lay reviewers. For each comparison, four kappas are shown: unweighted (for partial agreement-unadjusted (for sampling probabilities), weighted-unadjusted, unweighted-adjusted, and weighted-adjusted. Kappas for comparisons between the neurologist and lay reviewers were comparable to those for comparisons between the lay reviewers. For the broad seizure classifications, generalized onset and partial onset, the unweighted-unadjusted kappas indicated that agreement was substantial (0.73 and 0.77) for the reviewer 1 comparison and moderate (0.53) and substantial (0.75) for the reviewer 2 comparison. Within partial-onset seizures, agreement was substantial for complex partial seizures (unweighted-unadjusted κ= 0.67, 0.79) but poor for simple partial seizures.

Within generalized-onset seizures, the unweighted-unadjusted kappas indicated that agreement was moderate-substantial for GTC, substantial-almost perfect for absence, moderate for myoclonic, and fair for atonic seizures.

Adjustment for sampling probabilities influenced the kappa values differently for different seizure types. For the reviewer 1 comparisons, the adjusted kappas were generally lower than the unadjusted kappas, with the exception of absence seizures. For the reviewer 2 comparisons, the relations between adjusted and unadjusted kappas fluctuated to a greater degree. Adjustment influenced the kappa values to a greater extent for rare than for common seizure types. Thus, kappas for myoclonic and atonic seizures decreased substantially as a result of the adjustment, while only small changes were observed in the broad categories of generalized- and partial-onset seizures and in primary or secondarily generalized tonicclonic (SGTC) seizures.

Based on the adjusted kappas, agreement for the broad categories of generalized- and partial-onset seizures was moderate for the reviewer 1 and moderate-almost perfect for the reviewer 2 comparisons. For absence and complex partial seizures, agreements were substantial-almost perfect for both pairwise comparisons (0.69 to 0.94), while for myoclonic and atonic seizures, agreements ranged from fair to poor (0.13 to 0.22). Adjusted kappas for simple partial seizures indicated poor agreement.

As with the comparisons between the lay reviewers, the weighted kappas (both adjusted and unadjusted) were higher than the unweighted kappas because they allowed for partial agreement; however, applying weights did not change the results substantially.

Discussion

Comparisons between the lay reviewers

Agreement between the lay reviewers was in the substantial-almost perfect range for all seizure types except simple partial seizures (table 3). In general, kappas for partial-onset seizures were somewhat higher than those for generalized-onset seizures, but the differences were small, and a similar pattern was observed for all three reviewer pairs.

Agreement between reviewers 1 and 2 was somewhat better than for the other two reviewer pairs. Although all three reviewers received the same amount of training, reviewers 1 and 2 worked together for a longer period of time and reviewed a larger number of subjects than did reviewer 3. We did not examine the degree to which reliability may have improved over time, but we suspect that participation in the consensus meetings provided continual training, thus improving reliability. Nonetheless, agreement was substantial for all three reviewer pairs, supporting reliability of our diagnostic method.

Although weighted kappas provide a more accurate assessment of agreement between reviewers than do unweighted kappas, selection of weights is arbitrary and may confuse the interpretation of results.[7] We classified cases as "possible" when there was some evidence for a specific seizure type but the evidence was not sufficient to make a definite diagnosis. Thus, arbitrary weights of 0.75 for "yes-versus-possible" and 0.25 for "possible-versus-no" were chosen a priori to reflect the ways in which reviewers classified ambiguous cases. With weighting, kappas increased in general, and the changes ranged from 1 to 31%, depending on the number of discordant possibles. When alternative weights of 0.5 were used for both yes-versus-possible and possible-versus-no, the results were virtually identical to those calculated previously.

Comparisons between the neurologist and lay reviewers

For most seizure types, adjusted-weighted kappas were comparable to weighted kappas for pairs of lay reviewers (table 3). Agreement was substantial-almost perfect for SGTC, complex partial seizures, any partial-onset seizure, GTC, and absence seizures, and moderate-substantial for any generalized-onset seizure. For myoclonic and atonic seizures, however, agreement was substantially lower in the neurologist-lay reviewer comparisons than in the comparisons between lay reviewer pairs.

The problems in classification of simple partial seizures reflect, at least in part, inherent difficulties in identifying simple partial seizures that occur *independently* (rather than as part of the same event) in patients with complex partial or secondarily generalized seizures. Reutens et al[8] also found lower reliability for simple partial than for other seizure types.

Difficulties with the design of our semistructured interview also contributed to disagreement in the diagnosis of both simple partial seizures and myoclonic seizures. Thus, in one section of the interview, we asked whether the subject's small seizures involved "sudden jerking of part or all of your body." This question was intended to ascertain myoclonic seizures, but many patients with either simple or complex partial seizures also answered "yes-" Since our original aim was to ascertain myoclonic seizures, we did not include sufficient follow-up questions to identify symptoms of focality or alteration in consciousness, in order to differentiate between myoclonic seizures and simple or complex partial seizures. (A subsequent section of the interview addressed these symptoms, but some subjects were not asked these questions.) We have revised the interview accordingly and anticipate improved reliability and validity with the new version.

All the patients included here were confirmed to have epilepsy. Thus we assessed agreement between reviewer pairs in the classification of seizures in patients with epilepsy rather than agreement in the diagnosis of epilepsy per se. In addition, because all the patients included were reviewed by a single neurologist, we did not assess variation among the three neurologists in terms of their agreement with the lay reviewers.

We are aware of three previous studies[3,8,9] that have examined reliability and validity of seizure diagnosis. These studies differed in design and hence tested different contributions to variability in seizure diagnosis. Bodensteiner et al[9] examined interrater reliability of seizure

classifications by pairs of neurologists, based on review of descriptions of children's seizures in medical records. In our previous study of this question,[3] we compared diagnoses made by a non-neurologist, based on data collected in semistructured interviews, with clinical diagnoses of neurologists with expertise in epilepsy. Reutens et al[8] compared diagnoses made by a neurologist, based on data in semistructured interviews, with clinical diagnoses made by a different neurologist. Finally, in the current study, we compared diagnoses made by lay reviewers, based on data in semistructured interviews, with each other and with those of a neurologist with expertise in epilepsy, based on the same data.

Reliability was lowest in the study by Bodensteiner et al,[9] which did not involve standardized methods for either collection or interpretation of data on seizure symptoms. Reliability was highest in the study by Reutens et al,[8] in which data were collected using standardized methods and were interpreted by persons with a high level of expertise. In our earlier study,[3] reliability was not much lower than that in the study by Reutens et al,[8] even though the reviewer was not a neurologist.

The current study addresses a different question: How consistent are different lay reviewers in interpreting interview data, and how similar is their interpretation to that of an expert neurologist? The results show a high level of consistency, providing further reassurance about our approach of using trained, nonexpert research assistants to interpret data collected in standardized interviews, using standardized diagnostic criteria.

Development and validation of standardized methods for diagnosis and classification are important not only for epilepsy but also for a wide range of neurologic disorders. Our experience illustrates that despite the complexities involved, methods can be developed for collecting valid, clinically detailed information on seizure disorders in large-scale epidemiologic studies. Efforts in this area will be enhanced by the establishment of widely accepted operational criteria for diagnosis and structured interview instruments for data collection. These instruments can also be useful in clinical settings, for teaching purposes, and for collection of data on seizure symptomatology by nonphysician clinic staff.

Acknowledgements

The authors are grateful to Drs. Bruce Link and Sharon Schwartz for advice on the statistical analyses.

Supported by NIH RO1-NS20656.

References

1. Commission on Classification and Terminology of the International League Against Epilepsy. Proposal for revised clinical and electroencephalographic classification of epileptic seizures. Epilepsia 1981;22:489–501. [PubMed: 6790275]
2. Ottman R, Susser M. Strategies for data collection in genetic epidemiology: the Epilepsy Family Study of Columbia University. J Clin Epidemiol 1992;45:721–727. [PubMed: 1619451]
3. Ottman R, Hauser WA, Stallone L. Semi-structured interview for seizure classification: agreement with physicians' diagnoses. Epilepsia 1990;31:110–115. [PubMed: 2406127]
4. Fleiss, JL. Statistical methods for rates and proportions. 2. New York: John Wiley; 1981.
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174. [PubMed: 843571]
6. Shrout PE, Spitzer RL, Fleiss JL. Quantification of agreement in psychiatric diagnosis revisited. Arch Gen Psychiatry 1987;44:172–177. [PubMed: 3813814]
7. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126:161–169. [PubMed: 3300279]
8. Reutens DC, Howell RA, Gebert KE, Berkovic SF. Validation of a questionnaire for clinical seizure diagnosis. Epilepsia 1992;33:1065–1071. [PubMed: 1464265]

9. Bodensteiner JR, Brownsworth RD, Knapik JR, Kanter MC, Cowan LD, Leviton A. Interobserver variability in the ILAE classification of seizures in childhood. Epilepsia 1988;29:123–128. [PubMed: 3258236]

Table 1

Weighted and unweighted kappa for seizure classification by lay reviewers

	Reviewer pair (no. of patients)								
	1–2 (N = 79)			1–3 (N = 57)			2–3 (N = 18)		
Seizure type	**N***	**Unweighted**	**Weighted†**	**N**	**Unweighted**	**Weighted**	**N**	**Unweighted**	**Weighted**
Partial onset									
SGTC	46	0.68	0.70	42	0.69	0.75	13	0.66	0.67
Complex partial	47	0.80	0.83	35	0.68	0.69	13	0.87	0.87
Simple partial	1	0.44	0.45	9	0.43	0.46	0	−0.15	−0.15
Any partial onset	56	0.86	0.86	46	0.64	0.64	15	1.00	1.00
Generalized onset									
GTC	21	0.67	0.74	8	0.60	0.66	2	0.73	0.93
Absence	9	0.79	0.86	7	0.55	0.61	1	0.64	0.64
Myoclonic	5	0.87	0.97	2	1.00	1.00	0	—	—
Atonic	2	0.66	0.72	2	0.59	0.77	0	—	—
Generalized nonconvulsive	14	0.89	0.89	9	0.65	0.65	1	0.64	0.64
Any generalized onset	21	0.77	0.77	10	0.59	0.59	2	0.46	0.46

SGTC Secondarily generalized tonic-clonic.

GTC Primary generalized tonic-clonic.

* Number of subjects classified as positive for each seizure type, based on the final diagnosis. Seizure types are not mutually exclusive.

† Weights calculated according to formulas given by Fleiss,[4] with “yes-versus-possible” = 0.75, and “possible-versus-no” = 0.25.

Table 2

mparisons of seizure diagnoses between neurologist and lay reviewers

	Neurologist versus reviewer 1 (N = 44)				Neurologist versus reviewer 2 (N = 32)				
*	Unweighted-unadjusted	Weighted-unadjusted†	Unweighted-adjusted‡	Weighted-adjusted§	N	Unweighted-unadjusted	Weighted-unadjusted	Unweighted-adjusted	Weighted-adjusted
9	0.78	0.81	0.62	0.71	13	0.56	0.61	0.73	0.76
3	0.79	0.80	0.74	0.75	9	0.67	0.71	0.69	0.78
	−0.13	−0.16	−0.03	−0.04	0	−0.07	−0.08	−0.15	−0.19
9	0.77	0.77	0.61	0.61	13	0.75	0.75	0.86	0.86
4	0.78	0.78	0.62	0.64	19	0.56	0.66	0.73	0.81
2	0.89	0.90	0.97	0.97	9	0.69	0.76	0.75	0.81
5	0.50	0.52	0.22	0.26	4	0.45	0.48	0.22	0.26
4	0.40	0.48	0.17	0.19	2	0.35	0.35	0.13	0.13
3	0.76	0.76	0.74	0.74	13	0.60	0.60	0.48	0.48
4	0.73	0.73	0.68	0.68	19	0.53	0.53	0.54	0.54

clonic.

.

tive for each seizure type, based on final diagnosis. Seizure types are not mutually exclusive.

t: 0.75 for “yes-versus-possible” and 0.25 for “possible-versus-no.” Not adjusted for sampling distribution.

lities of generalized and partial onset seizures. Not weighted.

t and adjusted for sampling probabilities.

Neurology. Author manuscript; available in PMC 2008 June 13.

Table 3

Summary of results for comparisons between lay reviewers and neurologist versus lay reviewers

	Pairs of lay reviewers (weighted kappa)			Neurologist versus lay reviewers (adjusted and weighted kappa)		
Seizure type	**Lowest**	**Highest**	**Description***	**Reviewer 1**	**Reviewer 2**	**Description**
Partial onset						
SGTC	0.67	0.75	Substantial	0.71	0.76	Substantial
Complex partial	0.69	0.87	Substantial	0.75	0.78	Substantial
Simple partial	−0.15	0.46	Fair-poor	−0.04	−0.19	Poor
Any partial onset	0.64	1.00	Substantial-almost perfect	0.61	0.86	Substantial-almost perfect
Generalized onset						
GTC	0.66	0.93	Substantial-almost perfect	0.64	0.81	Substantial-almost perfect
Absence	0.61	0.86	Substantial-almost perfect	0.97	0.81	Substantial-almost perfect
Myoclonic	0.97	1.00	Almost perfect	0.26	0.26	Fair
Atonic	0.72	0.77	Substantial	0.19	0.13	Slight
Generalized nonconvulsive	0.64	0.89	Substantial-almost perfect	0.74	0.48	Moderate-substantial
Any generalized onset	0.46	0.77	Moderate-substantial	0.68	0.54	Moderate-substantial

SGTC Secondarily generalized tonic-clonic.

GTC Primary generalized tonic-clonic.

* Adjectives suggested by Landis and Koch.[5]