An Exploration of the Diagnostic Reliability of EEG Using Latent Class Analysis.

Abdel Baki SG[1], Weedon J[4], Arnedo V[2], Chari G[2,3], Koziorynska E[2], Lushbough CA[2], Maus D[2], McSween T[2,3], Mortati KA[2], Omurtag A[1], Reznikov A[2,3], Grant AC[2].

[1]Bio-Signal Group, Brooklyn, NY, and Departments of [2]Neurology, [3]Pediatrics, and [4]Scientific Computing Center, SUNY Downstate Medical Center and KCHC, Brooklyn, NY

INTRODUCTION

▪Inter-rater reliability (IRR) of EEG interpretation among 6 board- certified clinical neurophysiologists (data reported separately) revealed that the level of agreement varied significantly with the "electrographic category" being assessed.

▪Previous studies have also shown that inter-rater agreement differs across EEG categories, with "superior" categories having high agreement among raters.

OBJECTIVES

▪Identify electrographic categories with a relatively high level of inter-rater agreement ("superior categories")

▪Use the Latent Class Model (LCM) to identify these categories.

◆LCM analysis is based on the speculation that there will be distinct groupings (classes) of EEG interpretations that have specific characteristics that influence IRR.

STUDY DESIGN / METHODS

EEGs

✧300 studies recorded from patients ≥1 year old
✧All identifying information removed except age and medications
✧Technologist comments during the recording not removed

Interpreters

✧ Six board-certified clinical neurophysiologists, 3 adult and 3 pediatric
✧ Interpreters (A-F) divided into 20 groups of 3 (ABC, ABD…DEF)
✧ Each interpreter belonged to 10 groups / each group read 15 EEGs
✧ Each reader interpreted 150 EEGs, each EEG interpreted by 3 readers

Interpretation Procedure

✧ Readers assigned each EEG to ≥ 1 of 7 categories - Normal (Nl), Epileptiform (Epi), Slowing (Slow), Epileptiform + Slowing (Epi+Slow), Seizure (Sz), Status Epilepticus (SE) and Uninterpretable (Ui).

✧Readers assigned a probability (subjective confidence) to each of the ≥ 1 selected categories, with the sum of all probabilities = 1.

RESULTS

3 EEG classes characterized by considerable agreement among raters

Class I (Normal): 30% of cases

EEG Category	Nl	SE	Sz	Epi+Slow	Epi	Slow	Ui
Reader A	55%	0%	4%	0%	7%	34%	0%
Reader B	69%	0%	7%	4%	0%	20%	0%
Reader C	84%	0%	3%	3%	3%	7%	0%
Reader D	80%	0%	3%	4%	0%	13%	0%
Reader E	97%	0%	0%	3%	0%	0%	0%
Reader F	81%	0%	4%	3%	0%	4%	7%

Class II (Slow): 44% of cases

EEG Category	Nl	SE	Sz	Epi+Slow	Epi	Slow	Ui
Reader A	2%	4%	2%	10%	0%	68%	14%
Reader B	0%	0%	0%	51%	0%	49%	0%
Reader C	4%	3%	0%	10%	0%	84%	0%
Reader D	26%	0%	0%	14%	0%	60%	0%
Reader E	29%	0%	0%	0%	0%	65%	5%
Reader F	9%	0%	0%	0%	0%	80%	11%

Class III (Epi+Slow): 26% of cases

EEG Category	Nl	SE	Sz	Epi+Slow	Epi	Slow	Ui
Reader A	0%	15%	0%	59%	19%	0%	6%
Reader B	0%	4%	7%	74%	15%	0%	0%
Reader C	0%	0%	4%	56%	36%	0%	4%
Reader D	6%	7%	4%	50%	29%	0%	4%
Reader E	6%	0%	4%	56%	34%	0%	0%
Reader F	4%	4%	0%	50%	27%	15%	0%

In the context of this study, LCM is based on the speculation that there may be different EEG categories or group of readings that have specific characteristics.

If this is true, the first question is: ***How many classes are there?*** For the data obtained in this study, the "right" answer seems to be 3:

- **Class I readings are usually classified by all readers as Normal**
- **Class II readings are generally agreed to be Slow**
- **Class III readings are typically agreed to contain Epi+Slow**

DISCUSSION

- **Class I readings: Estimated to comprise 30% of the sampled population, these studies were most likely to be classified as Normal. However , Reader A, and to a lesser extent Reader B, were less likely to do this than the others, often interpreting these studies as Slow.**
- **Class II readings: Estimated to comprise 40% of the sampled population, were generally agreed to be Slow, though Readers D & E often categorized them as Nl, and Reader B as Slow+Epi.**
- **Class III readings: Estimated to comprise 26% of the sampled population, were typically agreed to contain Epi+Slow, though there is considerable disagreement about the Epi component.**

CONCLUSIONS

▪Based on these data obtained from 6 readers and 300 EEGs, the LCM resulted in 3 EEG classes characterized by considerable agreement among raters.

▪This finding is consistent with the concept of "superior" EEG features formulated by Abend NS et al, which relates inter-rater agreement to the characteristics of the EEG category being assessed.

▪EEG diagnostic categories with relatively little agreement among the raters should be a priority for continuing education of EEG intrepreters

Support NIH grant RC3NS070658-01. The authors are grateful to Persyst Development Corporation for providing the Insight reading software to the readers.

Selected References

1. Walczak TS et al. Accuracy and interobserver reliability of scalp ictal EEG. Neurology 1992;42:2279-85.
2. Abend NS., et al. Interobserver reproducibility of electroencephalogram interpretation in critically ill children. J ClinNeurophysiol. 2011; 28:15-9
3. Gerber PA., et al. Interobserver agreement in the interpretation of EEG patterns in critically ill adults. J ClinNeurophysiol. 2008; 25:241-9
4. Azuma H, et al. An intervention to improve the interrater reliability of clinical EEG interpretation. Psych Clin Neurosci. 2003; 57:485-489

Inter-rater Reliability (IRR) of EEG Interpretation: A Large Single-Center Study.

Chari G[1,2], Abdel Baki SG[4], Arnedo V[1], Koziorynska E[1], Lushbough CA[1], Maus D[1], McSween T[1,2], Mortati KA[1], Omurtag A[4], Reznikov A[1,2], Weedon J[3], Grant AC[1].

Departments of [1]Neurology, [2]Pediatrics, and [3]Scientific Computing Center, SUNY Downstate Medical Center and KCHC, Brooklyn, NY, and [4]Bio-Signal Group, Brooklyn, NY

INTRODUCTION

- **Inter-rater reliability (IRR) of EEG interpretation has significant clinical implications, since there is no gold standard for an EEG's "true" interpretation.**
- **Despite a standardized terminology for EEG findings, EEG interpretation remains inherently subjective.**
- **Previous studies of IRR have been limited by the number of interpreters, number of EEGs, heterogeneity of EEGs, and methods of categorizing EEG findings.**

OBJECTIVES

- **Examine IRR among 6 board-certified raters interpreting 300 EEGs using:**
  - **Cohen Kappa coefficient as a measure of agreement among rater pairs**
  - **Fleiss Kappa coefficient as a measure of agreement among rater groups and all raters (see below)**

STUDY DESIGN / METHODS

EEGs

- **300 studies recorded from patients ≥1 year old**
- **All identifying information removed except age and medications**
- **Technologist comments during the recording not removed**

Interpreters

- **Six board-certified clinical neurophysiologists, 3 adult and 3 pediatric**
- **Interpreters (A-F) divided into 20 groups of 3 (ABC, ABD…DEF)**
- **Each interpreter belonged to 10 groups / each group read 15 EEGs**
- **Each reader interpreted 150 EEGs, each EEG interpreted by 3 readers**

Interpretation Procedure - *Level 1*:

- **Readers assigned each EEG to ≥ 1 of 7 categories - Normal (Nl), Epileptiform (Epi), Slowing (Slow), Epileptiform + Slowing (Epi+Slow), Seizure (Sz), Status Epilepticus (SE) and Uninterpretable (UI).**
- **Readers assigned a probability (subjective confidence) to each of the ≥ 1 selected categories, with the sum of all probabilities = 1.**
- **One category had to have the highest probability. *Only data from the category with the highest probability are presented here.***

Interpretation Procedure - *Level 2*:

- **If appropriate for the level 1 category assigned the highest probability, readers further qualified the abnormality as focal, generalized, multifocal, or focal + generalized. *These data are not presented here.***

Statistical Methods

- **Fleiss Kappa (K_f) used to measure the IRR among all 6 readers.**
- **Cohen's Kappa coefficient (K_c) used to measure IRR among rater pairs, and was compared to the aggregate K_f.**
- **The strength of agreement using kappa values was determined as described by Landis and Koch.**

KAPPA INDEX stratified by Landis and Koch	
Kappa coefficient	**Classification**
< 0.20	Poor
0.21 to 0.40	Weak
0.41 to 0.60	Moderate
0.61 to 0.80	Good
> 0.81	Very Good

RESULTS

Table 1: Frequency of categorical assessment for each rater.
Note that no two raters interpreted the same set of 150 EEGs.

Reader	Overall							
	Epi	Slow	Epi+ Slow	SE	Sz	Nl	UI	Total
A	8	59	33	8	3	27	12	150
B	6	38	63	1	7	35	0	150
C	15	54	26	1	2	50	2	150
D	12	40	35	3	5	52	3	150
E	16	38	25	0	2	67	2	150
F	13	65	16	1	1	47	7	150
Total	70	294	198	14	20	278	26	900

Table 2. Frequency of categorical assessment for 3 condensed categories.

Reader	Table of Reader by Overall (3 Categories)			
	Epi/Slow/Epi+Slow	SE/Sz	Nl	Total
A	100	11	27	138
B	107	8	35	150
C	95	3	50	148
D	87	8	52	147
E	79	2	67	148
F	94	2	47	143
Total	562	34	278	874

- **For all 7 EEG interpretation categories, IRR agreement was:**
  1. **Moderate for Nl, Epi, and Slow**
  2. **Weak for Epi+Slow, SE, and Sz**
- **For the 3 condensed categories agreement was:**
  1. **Moderate for SE/Sz combined**
  2. **Moderate for Epi/Slow/Epi+Slow combined**

Table 3: IRR agreement for each category, aggregated among the 7 categories

Category	Fleiss Kappa Statistic	*P*-value (H_0 : no agreement)
Epi+Slow	0.391	<.001
Epi	0.458	<.001
Slow	0.459	<.001
Nl	0.558	<.001
SE	0.274	<.001
Sz	0.284	<.001
Ui	0.129	<.001
Aggregated	0.450	<.001

Table 4: IRR agreement for a 3 category classification

Category	Fleiss Kappa Statistic	*P*-value (H_0 : no agreement)
Nl	0.575	<.001
Epi/Slow/Epi+Slow	0.521	<.001
SE/Sz	0.417	<.001
Aggregated	0.537	<.001

The IRR agreement as measured by Fleiss kappa coefficients for EEG classes were as follows:

- **The aggregate agreement over all 7 categories was in the moderate range (Fleiss Kappa statistic: 0.45), consistent with a Cohen Kappa statistic of 0.43, when rater pair agreement was summarized across all EEG categories (Table 3).**
- **The agreement for ictal EEGs (SE/Sz) combined was moderate (0.417), with a better agreement for combined Epi/Slow/Epi+Slow (0.52) (Table 4).**

DISCUSSION

- **This is the first study of EEG inter-rater reliability to:**
  - ✓ **Use a large number of board-certified raters (6)**
  - ✓ **Use a large number of EEGs (300)**
  - ✓ **Require raters to assign each EEG to ≥ 1 of 7 categories**
  - ✓ **Require raters to assign a probability to each chosen category**
  - ✓ **Determine kappa values based on category of EEG interpretation**
- **The results highlight the variability of EEG interpretation among expert readers at one institution.**
- **None of the Fleiss kappa scores exceeded 0.6**
- **Inter-rater agreement was highest (Kappa=0.56) for the Normal category, and second highest (Kappa=0.46) for both the Epi and Slow categories.**
- **Low agreement for SE and Sz categories represents a potentially serious limitation in EEG interpretation.**

CONCLUSIONS

- **Inter-rater reliability of EEG interpretation is "moderate," even among expert readers.**
- **IRR may be improved by:**
  - **Establishing consensus guidelines for defining EEG features**
  - **Review of controversial EEG findings among groups of EEG readers to arrive at a consensus interpretation**
  - **Dissemination of consensus guidelines and principles to students and teachers of EEG interpretation**
- **A web-based case manager network allowing for the mitigation of IR variability by modifying the interpretation process across multiple interpreters at multiple institutions would be a strategy to achieve these goals.**

Selected References

1.Walczak TS et al. Accuracy and interobserver reliability of scalp ictal EEG. Neurology 1992;42:2279-85.

1.Abend NS., et al. Interobserver reproducibility of electroencephalogram interpretation in critically ill children. J ClinNeurophysiol. 2011; 28:15-9

1.Gerber PA., et al. Interobserver agreement in the interpretation of EEG patterns in critically ill adults. J ClinNeurophysiol. 2008; 25:241-9

1.Azuma H, et al. An intervention to improve the interrater reliability of clinical EEG interpretation. Psych Clin Neurosci. 2003; 57:485-489