Inter- and Intra-rater Reliability (IIRR) of EEG Interpretation: a large, single-center study.

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

[1] Bio-Signal Group, Brooklyn, NY; [2]Scientific Computing Center, Downstate Medical Center (DMC), SUNY, and KCHC Brooklyn, NY ; [3] Neurology, DMC, SUNY, and KCHC Brooklyn, NY; [4]Neurology and Pediatrics, DMC, SUNY, and KCHC Brooklyn, NY;

INTRODUCTION

INTRODUCTION

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

OBJECTIVES

- Examine IIRR 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 patient age and medications
- Technologist comments during the recording were 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 group / each group read 15 EEGs
- 300 EEGs were divided between two interpretation time intervals (T1 and T2), separated by several months.
- Each EEG was interpreted by 3 readers.
- Each reader interpreted 100 EEGs at both T1 and T2, 150 of which were unique
- EEGers were unaware that 50 EEGs per reader interpreted at T1 were reinterpreted at T2.

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
- One category had to have the highest probability

Statistical Methods

- Fleiss kappa (Kf) used to measure the Inter-rater reliability among all 6 readers
- Cohen's Kappa coefficient (Kc) used to measure Inter-rater reliability among rater pairs and within rater agreement (Intra-rater reliability).

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.

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

Table 2: Inter-rater 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

- For all 7 EEG interpretation categories, Inter-rater agreement was:
  1. Moderate for Nl, Epi, and Slow
  2. Weak for Epi+Slow, SE, and Sz

Table 3: Intra-rater assessment. Frequency of electrographic categories for each rater (A-F) in T1 and T2.

Frequency of categorical assessment for each rater in T1 and T2						
Scorer	A	B	C	D	E	F
	(T1) (T2)					
Epi+Slow	(7) (7)	(19) (13)	(10) (8)	(8) (10)	(7) (8)	(6) (9)
Epi	(6) (2)	(4) (4)	(10) (8)	(8) (5)	(9) (9)	(7) (4)
Slow	(15) (20)	(10) (13)	(11) (12)	(12) (10)	(13) (13)	(15) (17)
Nl	(6) (12)	(12) (16)	(15) (20)	(17) (18)	(18) (19)	(13) (12)
SE	(6) (2)	(1) (1)	(1) (0)	(2) (1)	(0) (0)	(1) (0)
Sz	(2) (2)	(4) (2)	(2) (1)	(2) (5)	(1) (1)	(1) (1)
UI	(8) (5)	(0) (1)	(1) (1)	(1) (1)	(2) (0)	(7) (7)

Table 4: Intra-rater agreement on 50 EEGs sample

Intra-observer agreement: kappa values (± SD) for 6 readers analyzing 50 EEG samples at the beginning of the study and several months later				
Scorers	Kappa	SD	95% Confidence Limits	
A	0.3314	0.0872	0.1604	0.5023
B	0.5016	0.0891	0.3270	0.6762
C	0.5801	0.0831	0.4172	0.7429
D	0.6651	0.0770	0.5143	0.8160
E	0.7299	0.0758	0.5813	0.8784
F	0.6423	0.0808	0.4839	0.8007
Aggregated	0.5883	0.0334	0.5229	0.6537

- Intra rater Kc ranged from 0.33 to 0.73 for individual readers, with an aggregated Kc of 0.59 (95%CI 0.523 0.653)

DISCUSSION

- This is the first study of EEG IIRR to:
  - ✓ Use a large number of board-certified raters
  - ✓ Use a large number of EEGs
  - ✓ 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
  - Low agreement for SE and Sz categories represents a potentially serious limitation in EEG interpretation
- Inter-rater reliability of EEG interpretation is "moderate" even among expert readers

CONCLUSIONS

- EEG inter-and intra-rater reliability are in the weak to good range in highly qualified and experienced EEGers
- The subjectivity of EEG interpretation may be reduced by:
  - ✓ Establishing consensus interpretation guidelines
  - ✓ Developing statistical prediction rules
  - ✓ Engaging neurologists in an on-line continuous review of challenging studies
- ➢ A web-based case manager network allowing for the mitigation of IIRR variability by modifying the interpretation process across multiple interpreters at multiple institutions would be a strategy to achieve these goals

Selected References

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

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