**On the Use of Non-Experts for   
Generation of High Quality Annotations of Seizure Events**

**Vinit Shaha, Eva von Weltina, Tameem Ahsana, Iyad Obeida and Joseph Piconea**

a The Neural Engineering Data Consortium, Temple University 1947 North 12th Street, Philadelphia, Pennsylvania, 19122, USA.  
{vinitshah, eva.vonweltin, tameem.ahsan, iobeid, picone}@temple.edu

**Abstract:** The ubiquity of electroencephalography (EEG) in the diagnosis of neurological disorders makes the automatic interpretation of EEG records, specifically with respect to the automatic detection of seizures, a prime target for the application of machine learning research. This type of research, however, requires an extremely large set (tens of thousands) of annotated EEG data to support the development of state of the art statistical models. Clinical interpretation of EEG recordings by certified neurologists is focused on diagnosis, not on the creation of detailed annotations. Engaging neurologists in the creation of seizure annotations with the level of detail necessary to conduct machine learning research is a slow, tedious and expensive process. This process is further complicated by inconsistent inter-rater agreement. In this paper, we demonstrate that it is possible to create a database of high quality seizure annotations using non-experts, undergraduate neuroscience students in this case, who are properly trained to interpret EEG waveforms. Inter-rater agreement between annotations made by neurologists and student annotators is evaluated using the Any-Overlap scoring method and using an Epoch-Based scoring method. Inter-rater agreement between neurologists and student annotators, using Cohen’s Kappa coefficient, is within the range of 0.53 - 1.00 indicating sufficiently strong agreement. This annotation team has successfully transcribed a large portion of the open source TUH EEG Seizure Detection Corpus, thereby enabling the development of high performance technology.

**Keywords:** Electroencephalography, EEG, Inter-rater agreement, machine learning

# Introduction

Electroencephalograms are the primary tool by which clinicians diagnose brain related illnesses such as epilepsy, non-epileptic seizures, and sleep disorders (Yamada and Meng, 2009). Seizures, which are seen most often in patients diagnosed with epilepsy, can occur in a convulsive or non-convulsive manner. In an ICU environment, approximately 90% of these seizures are clinically unrecognizable non-convulsive seizures which can only be diagnosed by continuous EEG (cEEG) monitoring (Hirsch, 2010). Though clinicians do periodically observe EEGs for the identification of such seizures, any delay in the treatment of non-convulsive seizures in ICU environments can be harmful and even deadly to patients (Hirsch and Kull, 2004; Wiebe, 2008). To aid in the speed and efficiency of the diagnosis and treatment process, automatic interpretation of EEGs has been studied over the past decade for application in clinically relevant software (Alotaiby et al., 2014; Gotman, 1982; Wilson et al., 2003). However, the development of such applications requires a large and, for many researchers, prohibitive amount of transcribed EEG data.

Annotation of EEGs is usually done by certified neurophysiologists who have received extensive training. In order to speed up the diagnosis process, experienced clinicians will rapidly skim through an EEG record and annotate any intervals in which interesting events occur using simple “start” and “stop” marks. As a result, clinicians will often miss some events, especially those that are subtle or brief. This process is thereby inclined to miss some, often subtle or very brief, events. Additionally, these transcriptions cannot be directly used for technological development due to a lack of detailed spatial information. This annotation process is subjective and relies on various clinical evidences including push button events and medication dosages. Due to inconsistencies in an individual’s judgement, poor inter-rater agreement (IRA) performance among neurologists is common on tasks such as detection of seizures and periodic discharges (PD) (Halford et al., 2015; Ronner et al., 2009).

Since the EEG is the primary tool used by neurologists for the diagnosis of neurological disorders, a significant portion of a typical neurologist’s professional life is assigned to EEG analysis. Considering that the average annual salary of a neurologist in the United States is approximately ~0.22 million US dollars, employing neurologists to create EEG annotations for research purposes can be prohibitively expensive. Furthermore, due to a lack of clear standards of interpretation and the creation of inexact annotations, this considerable effort can unfortunately result in inconsistent and unreliable data. In this study, we show that it is possible to develop a large, standardized, and annotated dataset by training undergraduate student annotators in the interpretation of deidentified EEGs. This process is faster, notably less expensive, and can result in superior inter-rater reliability than can be seen by employing neurologists in the development of a dataset of comparable size.

Publicly available annotated EEG databases are scarce and under-representative of the diverse population of patients seen in real world clinical settings. For example, one of the most prominent databases available is CHB-MIT which contains only 23 subjects (Goldberger et al., 2000). Larger databases made possible by rapid and precise annotators would immensely help research efforts. A rapidly evolving, standardized, and high-volume set of EEG annotations would eventually allow the fields of machine learning (ML) and artificial intelligence (AI) to move forward at a faster pace than previously imagined.

The subsets used for this study are randomly sampled from datasets collected at Duke University, Emory University, both of which were collected for the study of qEEG, and from the Temple University Hospital EEG Seizure corpus (Haider et al., 2016; Swisher et al., 2015;Shah et al., 2018). These three sets were independently evaluated because the modalities in which these databases were collected such as sources, database collection period, specific neurologists who defined “gold-standard” reference annotations, purpose of study, annotator experience, etc. varied with each set. In this study, we evaluate the performance of trained undergraduate student annotators on identification of seizure events that have been annotated by trained neurologists, as well as the evaluation of inter-rater agreement (IRA) performance on all the datasets. An example of a neurologist’s detailed annotations that we collected for our Temple University IRA test is shown in . For the sake of comparison, we define “gold-standard” annotations as those created by a group of neurologists and “aggregate-standard” as annotations created by a group of undergraduate students. Gold-standard annotations are defined by the agreement among two or more neurologists. Individual expert annotators’ annotations are not considered as a standard. Aggregate-standard annotations have been defined by conducting group meetings and discussions to agree and establish consent on marked annotations by the student annotators.

# Method

A close up of a map

Description generated with high confidence

Figure 1 Example of Neurologists’ descriptive annotations of TUSZ-IRA test set

## Subject population and test EEG dataset

The records used in this study are collected from three EEG seizure data sources from three different medical institutions. The subset TUSZ-IRA contains pruned records from 5 subjects with a total duration of 25,940 seconds. The DUSZ-IRA and EUSZ-IRA datasets both contain continuous EEG (cEEG) records collected from 5 patients, consisting of 72,001 seconds of data, and 3 patients consisting of 66,530 seconds of data respectively, both of which are comprised of critically ill ICU patients. The collection of these EEG data was approved by the individual institutions as well as XXXX board. Each inter-rater agreement test was performed independently such that trained student annotators were provided a combination of ictal and non-ictal files in TUSZ-IRA and ictal-only cEEG files from DUSZ-IRA and EUSZ-IRA. The gold-standard annotations of the DUSZ-IRA and EUSZ-IRA subsets were generated by neurologists at these respective institutions. In the DUSZ-IRA set these gold-standard annotations represent an agreement met by two neurologists (Swisher et al., 2015). In the EUSZ-IRA set these gold-standard annotations were annotated independently by three different neurologists (Haider et al., 2016). The TUSZ-IRA was originally distributed to 23 neurologists, four of whom completed and returned their annotations within the suggested timeline. According to gold-standard seizure annotations, there were 157 seizures of diverse focality, morphology, and duration between all IRA subsets. The distribution of number of seizures based on their duration are shown in . The seizures collected from TUSZ were within the range of 1 to 5 minutes. The majority of seizures in the DUSZ distribution have a duration within 30 seconds to 3 minutes. The majority of EUSZ seizures were 1 to 3 minutes long.

## Expert raters

A close up of a logo

Description generated with high confidence

Figure 2 Seizure duration distribution of TUSZ, DUSZ and DUSZ IRA subsets

All three sets were annotated by a group of 4 or 5 student employees who underwent 2-3 months of extensive training in the interpretation of EEGs and in the precise annotation of seizure events. This is in contrast with the annotations collected from board-certified EEG experts who have experience reviewing intensive care unit (ICU) EEGs rather than annotation for research purposes. The IRA tests were bi-class (ictal vs. non-ictal) classification processes where all the experts were asked to simply mark onset and offset of the ictal events. No consideration was made to the start and end times of individual channels. The evaluation is made on intra-expert (within student/neurologist annotators) and inter-expert (gold-standard Vs. aggregated-standard annotations) level. A total of 14 cEEG files (5 + 9 from DU and EU) and 32 pruned files (from TU) were used for three sessions of IRA tests which, according to gold-standard annotations, contained total seizure duration of 13,054 seconds (out of total 164,471 seconds). Though each annotator reviewed the EEG records independently, student annotators were allowed use books, notes and web resources as general references.

## Evaluation metrics and inter-rater agreement analysis

Relative similarity between independently generated annotations can be evaluated using the Kappa statistic. We have used Cohen’s kappa coefficient as our inter-rater agreement analysis tool. Cohen’s Kappa coefficients were calculated for each pair of raters at an intra-expert level. The same statistic was used to evaluate IRA of gold-standard and aggregate-standard annotations at an inter-expert level. The Cohen’s kappa coefficient can be calculated as:

where is the relative observed agreement between raters (observed accuracy) and  is the hypothetical probability of chance agreement (expected accuracy). Values below 0 suggest no agreement, 0-0.20 as slight agreement, 0.21-0.40 as fair agreement, 0.41-0.60 as moderate agreement, 0.61-080 as substantial agreement and 0.81-1.0 suggests almost complete agreement (Landis and Koch, 1977).

Intermediate variables in the calculation of Cohen’s Kappa such as observed accuracy and expected accuracy that are necessary for IRA analysis can be computed from the parameters true positives, true negatives, false positives and false negatives collected from the scoring methods used. We have collected these parameters from two scoring metrics: Any-overlap (OVLP) scoring and epoch based scoring (EPOCH) (Ziyabari et al., 2017). The OVLP method considers a detection as any case of overlap between events. In the EPOCH scoring metric, a record gets divided into equally sized subsamples called an “epoch” (defined as 1 second for our analysis) and scoring is performed on each epoch independently. compares the functionality of the OVLP and EPOCH methods. The Kappa coefficient is calculated using each of these metrics and is cross-tabulated to represent intra-expert and inter-expert analysis for all three seizure sets.

A screenshot of a cell phone

Description generated with high confidence

A close up of a map

Description generated with high confidence

Figure Functionality of Overlap Scoring metric (Top) and Epoch scoring metric (Bottom)

Identification of exact ictal onset and offset can be difficult in some ictal events, especially those which evolve very slowly. If the majority of a hypothesis annotation overlaps temporally with a reference annotation, give or take a few seconds at the very beginning and very end of the event, then this should be considered as a detection. The OVLP addresses this concern by considering any overlap as a detection. This is in contrast to EPOCH scoring which evaluates agreement at every second. Though EPOCH scoring is not so forgiving of raters over or under annotating events, we have seen that, in comparisons of prolonged events with higher true detection scores (TP and TN), these detection scores will not be overwhelmed by the error scores (FP and FN).

A screenshot of a cell phone

Description generated with very high confidence

A screenshot of a cell phone

Description generated with very high confidence

Figure 4 An example of inconsistent parameters generated by OVLP scoring method

It is worth noting that the EPOCH metric is biased towards longer events. So, annotating the longest seizures will improve the results, but at the same time, early or late onset and offset annotations for events can deteriorate the results. Though kappa statistic tables, including those created using the EPOCH method, are often symmetric, this is not the case when calculated using parameters generated by the OVLP method. presents an example of a seizure event in which the kappa statistic calculated from parameters generated using the OVLP method will be different depending upon which rater is taken to be the reference. In the first case of rater A versus rater B, 3 hypotheses annotations overlap with the first reference annotation, resulting in a single positive detection on that event. In the alternate case with rater B versus rater A shows that a single hypothesis annotation overlaps with 3 reference annotations, resulting in 3 positive detections. For this reason, we always include full matrices when reporting kappa statistics, not just one diagonal.

# Results

## Temple University IRA set (TUSZ-IRA)

From the TUH EEG Seizure Detection Corpus, 5 patients containing 32 pruned (<1 hour long) files were selected with a total duration of 25,918 seconds for the TUSZ-IRA test. According to the most recent version of TUSZ (v1.2.0), there are 12 seizures in this subset. First, we consider the intra-expert neurologist’s annotations that were collected for the TUSZ-IRA subset. This test set was assigned to 4 neurologists. From 4 of these neurologists, one pair was assigned 14 identical files and another pair was assigned 18 identical files. Intra-expert level agreement of two pairs are shown in . The agreement between clinician annotator 1 (CAnn-1) and clinician annotator 2 (CAnn-2) on 14 files is almost perfect (~1) according to both OVLP and EPOCH methods. Similar evaluation on clinician annotator 3 (CAnn-3) and clinician annotator 4 (CAnn-4) shows very poor performance (~0.2) according to both IRA metrics. This is a result of CAnn-4 marking seizures parsimoniously whereas CAnn-3 marked seizures very generously.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Overlap | | hyp | |  | Overlap | | hyp | |
| CAnn-1 | CAnn-2 | CAnn-3 | CAnn-4 |
| ref | CAnn-1 | 1.000 | 1.000 | ref | CAnn-3 | 1.000 | 0.2 |
| CAnn-2 | 1.000 | 1.000 | CAnn-4 | 0.2 | 1.000 |
| Epoch | | hyp | | Epoch | | hyp | |
| CAnn-1 | CAnn-2 | CAnn-1 | CAnn-1 |
| ref | CAnn-1 | 1.000 | 0.98 | ref | CAnn-3 | 1.000 | 0.159 |
| CAnn-2 | 0.98 | 1.000 | CAnn-4 | 0.159 | 1.000 |

Table 1 Neurologist’s pairwise agreements on TUSZ IRA subset

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Overlap | | hypothesis | | | | | | |
| Gold-St | Agg-St | StAnn1 | StAnn2 | StAnn3 | StAnn4 | StAnn5 |
| reference | Gold-St | 1.000 | 1.000 | 0.949 | 0.837 | 0.815 | 0.945 | 0.949 |
| Agg-St | 1.000 | 1.000 | 0.949 | 0.837 | 0.815 | 0.945 | 0.949 |
| StAnn1 | 0.949 | 0.949 | 1.000 | 0.895 | 0.865 | 0.895 | 1.000 |
| StAnn2 | 0.836 | 0.836 | 0.894 | 1.000 | 0.762 | 0.886 | 0.894 |
| StAnn3 | 0.818 | 0.818 | 0.867 | 0.767 | 1.000 | 0.767 | 0.867 |
| StAnn4 | 0.945 | 0.945 | 0.894 | 0.886 | 0.762 | 1.000 | 0.894 |
| StAnn5 | 0.949 | 0.949 | 1.000 | 0.895 | 0.865 | 0.895 | 1.000 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Epoch | | hypothesis | | | | | | |
| Gold-St | Agg-St | StAnn1 | StAnn2 | StAnn3 | StAnn4 | StAnn5 |
| reference | Gold-St | 1.000 | 0.870 | 0.829 | 0.727 | 0.881 | 0.831 | 0.886 |
| Agg-St | 0.870 | 1.000 | 0.812 | 0.828 | 0.933 | 0.846 | 0.932 |
| StAnn1 | 0.829 | 0.812 | 1.000 | 0.764 | 0.856 | 0.824 | 0.869 |
| StAnn2 | 0.727 | 0.828 | 0.764 | 1.000 | 0.835 | 0.858 | 0.832 |
| StAnn3 | 0.881 | 0.933 | 0.856 | 0.835 | 1.000 | 0.869 | 0.959 |
| StAnn4 | 0.831 | 0.846 | 0.824 | 0.858 | 0.869 | 1.000 | 0.865 |
| StAnn5 | 0.886 | 0.932 | 0.869 | 0.832 | 0.959 | 0.865 | 1.000 |

Table 2 TUSZ-IRA test’s pairwise kappa value between student annotators and gold-standard annotations

TUSZ-IRA gold-standard annotations were established by taking aggregate markings of the intersecting annotations of neurologists for the first group. This approach wasn’t pragmatic for the second group due to a low IRA. Here, clinician annotator CAnn-3’s annotations were taken as the ground-truth after review of these annotations by an independent expert.

shows pairwise performance on the inter-expert level as gold-standard (Gold-St) vs. aggregate standard (Agg-St) as well on the intra-expert level for student annotators (StAnn). The Gold-St vs. Agg-St pair shows almost perfect agreement between two groups with kappa values at 1.0 according to OVLP and 0.87 according to EPOCH. On the intra-expert level, every StAnn pair’s IRA scores according to OVLP and EPOCH kappa values show substantial to perfect agreement. From the OVLP IRA results, it can be seen that StAnn1 and StAnn5 show perfect agreement on individual seizure events. Less agreement between StAnn1 and StAnn2 indicates that they have disagreement seizure/non seizure decisions as well as the duration of seizures. Overall, the TUSZ-IRA test has highest agreement between among all the IRA tests conducted in this study.

## Duke University IRA test (DUSZ-IRA)

The Duke university IRA test-set contained records related to 5 patients and 5 ICU-CEEG files for a total duration of 72,001 seconds. According to received gold-standard annotations from Duke University, these files contain 63 seizures for a total seizure duration of 5202 seconds. These gold-standard neurologist’s annotations were compared to 4 student annotators’ annotations. Shows the pairwise agreement between individual student annotators, aggregate student annotations, and gold-standard annotations.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Overlap | | hypothesis | | | | | |
| Gold-St | Agg-St | StAnn1 | StAnn2 | StAnn3 | StAnn4 |
| reference | Gold-St | 1.000 | 0.729 | 0.729 | 0.327 | 0.419 | 0.766 |
| Agg-St | 0.745 | 1.000 | 0.974 | 0.458 | 0.556 | 0.862 |
| StAnn1 | 0.745 | 0.974 | 1.000 | 0.487 | 0.586 | 0.862 |
| StAnn2 | 0.428 | 0.536 | 0.570 | 1.000 | 0.582 | 0.556 |
| StAnn3 | 0.468 | 0.575 | 0.605 | 0.512 | 1.000 | 0.574 |
| StAnn4 | 0.780 | 0.861 | 0.861 | 0.475 | 0.555 | 1.000 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Epoch | | hypothesis | | | | | |
| Gold-St | Agg-St | StAnn1 | StAnn2 | StAnn3 | StAnn4 |
| reference | Gold-St | 1.000 | 0.736 | 0.686 | 0.531 | 0.576 | 0.708 |
| Agg-St | 0.736 | 1.000 | 0.921 | 0.627 | 0.679 | 0.833 |
| StAnn1 | 0.686 | 0.921 | 1.000 | 0.641 | 0.664 | 0.826 |
| StAnn2 | 0.531 | 0.627 | 0.641 | 1.000 | 0.657 | 0.616 |
| StAnn3 | 0.576 | 0.679 | 0.664 | 0.657 | 1.000 | 0.647 |
| StAnn4 | 0.708 | 0.833 | 0.826 | 0.616 | 0.647 | 1.000 |

Table 3 Pairwise comparison between Student annotators and gold-standard annotations on Duke IRA subset

The IRA kappa value between individual student annotators and gold-standard annotations collected from DU range from moderate to substantial according to both scoring metrics where StAnn1 and StAnn4 shows maximum agreement within the group as well as compared to gold-standard annotations. OVLP results generated DUSZ-IRA test has a very non-symmetric matrix which is representative of the example shown in . StAnn-2 and StAnn-3 are doing poorly according to both gold-standard and aggregate-standard annotations.

## Emory University IRA test (EUSZ-IRA)

The final subset in this study contained 9 ICU-CEEG files (duration: 66,530 seconds) related to 3 patients collected from Emory University Hospital. According to gold-standard annotations received from this institution, this subset contains 82 seizure events with a total duration of 5870 seconds. Gold-standard annotations of this test were agreed upon by at least three neurologists. shows the pairwise agreement between individual annotator pairs, aggregate-standard and gold-standard annotations.

IRA between two groups of experts on the EUSZ-IRA test was substantial to almost perfect. The majority of the annotations created by student annotators are in nearly perfect agreement with respect to both gold and aggregate standards.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Overlap | | hypothesis | | | | | |
| Gold-St | Agg-St | StAnn1 | StAnn2 | StAnn3 | StAnn4 |
| reference | Gold-St | 1.000 | 0.834 | 0.798 | 0.833 | 0.823 | 0.852 |
| Agg-St | 0.838 | 1.000 | 0.923 | 0.882 | 0.967 | 0.912 |
| StAnn1 | 0.796 | 0.920 | 1.000 | 0.865 | 0.909 | 0.895 |
| StAnn2 | 0.838 | 0.882 | 0.869 | 1.000 | 0.882 | 0.890 |
| StAnn3 | 0.837 | 0.969 | 0.917 | 0.888 | 1.000 | 0.906 |
| StAnn4 | 0.849 | 0.908 | 0.894 | 0.884 | 0.896 | 1.000 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Epoch | | hypothesis | | | | | |
| Gold-St | Agg-St | StAnn1 | StAnn2 | StAnn3 | StAnn4 |
| reference | Gold-St | 1.000 | 0.779 | 0.752 | 0.740 | 0.739 | 0.758 |
| Agg-St | 0.779 | 1.000 | 0.967 | 0.825 | 0.825 | 0.855 |
| StAnn1 | 0.752 | 0.967 | 1.000 | 0.818 | 0.827 | 0.836 |
| StAnn2 | 0.740 | 0.825 | 0.818 | 1.000 | 0.823 | 0.853 |
| StAnn3 | 0.739 | 0.825 | 0.827 | 0.823 | 1.000 | 0.848 |
|  | StAnn4 | 0.758 | 0.855 | 0.836 | 0.853 | 0.848 | 1.000 |

Table 4 Pairwise comparison between student annotators and gold-standard annotations on EUSZ-IRA subset

## Statistical Inference

For statistical inference, direct comparison of supersets containing all gold-standard annotations to aggregate or individual annotator pairs are considered. All tests are performed based on the Epoch scoring metric with an epoch duration of 1 second. (top left) shows the histogram of kappa scores based on individual files. To understand the normality of the distribution we performed a one-sample Kolmogorov-Smirnov test which yields a KS value of 0.685 (p-value < 0.001), indicating that the distribution can be considered as Gaussian. This same distribution is shown as a boxplot in the top right corner of the figure. It can be seen that overall IRA on most files is in the substantial to perfect range with some outliers in the moderate range. Whiskers of this plot spread from 0.62 in the lower range to 0.96 in the upper range. The bottom left figure shows individual boxplots for each IRA subset. Here, the second and third quartiles are in almost perfect agreement range for TUSZ and EUSZ. The DUSZ distribution has more range with its median value around ~0.7 and its lower whisker spreading to 0.5.

Performing a one-way ANOVA test on all student annotators rejects the null hypothesis with an F-value of 1.42 with p-value 0.239 and an F-value of 0.49 with p-value 0.684 for sensitivity and specificity of EPOCH scoring metric respectively. From (bottom right), it can be seen that Pearson’s correlation coefficient calculated on time series for individual annotator’s annotations are highly correlated with respect to gold-standard annotations (p-value < 0.001). Both tests suggest that student annotators’ performance is very similar to each other as no one outperforms any other rater when compared to the gold-standard. These consistent results are an important factor in justifying the development standard annotations. The correlation between the two IRA metrics being used, OVLP and EPOCH, are correlated with a ρ-value of 0.58 and a p-value of 0.002.

# Discussion

A screenshot of a cell phone

Description generated with very high confidenceA screenshot of a cell phone

Description generated with very high confidence

A screenshot of a cell phone

Description generated with very high confidenceA screenshot of a cell phone

Description generated with very high confidence

Figure 5 Student annotators IRA results comparison with neurologists’ gold standard annotations based on EPOCH scoring metric

Continuous EEG (cEEG), which contains hours to multiple days’ worth of data, is becoming a vital tool for correctly diagnosing patients with epilepsy and other brain related diseases. Hours’ worth of recordings for each patient under a neurologist’s care amounts to a very large volume of data that then needs to be evaluated for diagnostic markers. Due to this high volume of data and the strain on their time, neurologists are accustomed to skimming quickly through the record for prominent events and clear state changes. This process, however, leads to the neurologists missing brief and/or low amplitude, yet still clinically relevant, events. Annotations created in this process are simple and cursory. They are not suited for use in a database intended for technology development. Instead, if trained students, especially those with an interest in neuroscience and neurology, are employed to make these transcriptions, the process can be improved in terms of speed, detail, consistency, and cost. This study also suggests that group discussions and meetings among student annotators can help in the establishment of standards for annotations and can aide in quality-control the annotated database.

The significance of this study can be seen in the hinging of decisions made by neurologists as to the diagnosis and severity of epilepsy in clinical settings upon the identification of seizures in a patient’s EEG record. Unfortunately, identification of seizures, even according to ACNS terminology, is very subjective and uses rules that cannot be applied universally such as whether frequency evolution of PLDs/GPDs appearing in long bursts, several post-status epilepticus stages, and low frequency (1-3 Hz) spike and wave discharges lasting for more than 10 seconds should be considered as seizures. These ambiguities can result in differing opinions even among top experts in the field. Moreover, even after consensus, establishment of agreement on onset/offset of these events is nearly impossible. Our annotation team at the Neural Engineering Data Consortium (NEDC) has been collecting and discussing difficult EEG records and sharing them online with experienced readers; this forum is open to the public and all are welcome to post their thoughts and analysis. This is a good way to establish consent on morphologies, which can tend towards biases when discussed among small groups. The link to our FAQ queries can be found at:  
https://www.isip.piconepress.com/projects/tuh\_eeg/faq/index.shtml

It is unclear why agreement between student annotators and neurologists was relatively low in the EUSZ-IRA subset. This could be attributed to experts considering several low amplitude morphologies as clinically irrelevant. Despite this, student annotators have managed to maintain consensus within the group (on intra-expert level). Compared to gold-standard annotations, student annotators’ mean sensitivity and specificity on all subsets are 80.77% and 97.14% respectively. Low specificity suggests that false alarms were avoided.

This study is more significant than some of previously conducted IRA tests in neurology due to following reasons:  
(1) All annotations were collected and annotated in a digital format; (2) The study was performed on routine EEG records (TUSZ) as well as cEEG records (DUSZ, EUSZ) from different institutions which is representative of real world clinical practices; (3) Instead of asking for absence or presence of ictal events in a particular segment, this study additionally emphasizes onset and offset of every single ictal event.(4) This study uses two separate scoring metrics, one on an event basis and other on an epoch basis, for detection of seizure events. This gives us two separate perspectives; specificity in terms of seizure duration, and base sensitivity to the event. (5) This study proposes a method for the continued and future generation of refined annotated data sources for use in the development of new clinical technology.

The tool our student annotators used to create and reviewing annotations is called “demo tool” (noah’s reference) which was developed by our team at NEDC. This EEG viewing and annotation tool (Capp et al., 2017) is now available at no cost to the public and can be found and downloaded at   
 https://www.isip.piconepress.com/projects/tuh\_eeg/downloads/nedc\_demo/

We had originally hoped to collect a large volume of annotations for our TUSZ-IRA test. As previously mentioned, we distributed our temple university seizure detection set (TUSZ) to more than 20 neurologists but only 4 of them could deliver us annotations. This shows the difficulty of engaging expert neurophysiologists in such studies and in general for detailed annotation process. The time commitment necessary to create precise, detailed annotations is simply not conducive to the busy schedule that neurologists contend with.

Finally, this study emphasizes the vital role that a large, detailed, and affordable dataset must play in the path forward of rapid technology development, and the usefulness of student annotators in this pursuit. It in no way attempts to question or challenge the abilities of expert neurophysiologists.

# Acknowledgment

The authors would like to thank Dr. Mercedes Jacobson of Temple University Hospital and Dr. Steven Tobochnik of New York Presbyterian Hospital at Columbia University for enabling the creation of this corpus and their instruction in the interpretation of clinical EEG data.

# References

Alotaiby T, Alshebeili S, Alshawi T, Ahmad I, Abd El-Samie F. EEG seizure detection and prediction algorithms: a survey. EURASIP J Adv Signal Process 2014;2014:1–21. doi:10.1186/1687-6180-2014-183.

Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Circulation 2000;101:e215–20.

Gotman J. Automatic recognition of epileptic seizures in the EEG. Electroencephalogr Clin Neurophysiol 1982;54:530–40. doi:https://doi.org/10.1016/0013-4694(82)90038-4.

Haider HA, Esteller R, Hahn CD, Westover MB, Halford JJ, Lee JW, et al. Sensitivity of quantitative EEG for seizure identification in the intensive care unit. Neurology 2016;87:935–44. doi:10.1212/WNL.0000000000003034.

Halford JJ, Shiau D, Desrochers JA, Kolls BJ, Dean BC, Waters CG, et al. Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings. Clin Neurophysiol 2015;126:1661–9. doi:10.1016/j.clinph.2014.11.008.

Hirsch LJ, Kull LL. Continuous EEG monitoring in the intensive care unit. Am J Electroneurodiagnostic Technol 2004;44:137–58. doi:10.1046/j.1528-1157.43.s.3.7.x.

Hirsch LJ. Urgent continuous EEG (cEEG) monitoring leads to changes in treatment in half of cases. Epilepsy Curr 2010;10:82–5. doi:10.1111/j.1535-7511.2010.01364.x.

Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics 1977;33:159. doi:10.2307/2529310.

Capp N, Krome E, Obeid I, Picone J. Rapid Annotation of Seizure Events Using an Extensible Visualization Tool. Proc. IEEE Signal Process. Med. Biol. Symp., Philadelphia, Pennsylvania, USA: 2017, p. 1. doi:https://www.isip.piconepress.com/publications/conference\_presentations/2017/ieee\_spmb/demo/.

Ronner HE, Ponten SC, Stam CJ, Uitdehaag BMJ. Inter-observer variability of the EEG diagnosis of seizures in comatose patients. Seizure 2009;18:257–63. doi:10.1016/j.seizure.2008.10.010.

Shah V, von Weltin E, Lopez S, McHugh JR, Veloso L, Golmohammadi M, et al. The TUH EEG Seizure Corpus. Frontiers in Neuroscience 2018:6. doi:https://www.isip.piconepress.com/publications/unpublished/journals/2018/frontiers\_neuroscience/tuh\_eeg\_seizure/.

Swisher CB, White CR, Mace BE, Dombrowski KE. Diagnostic Accuracy of Electrographic Seizure Detection by Neurophysiologists and Non-Neurophysiologists in the Adult ICU Using a Panel of Quantitative EEG Trends. J Clin Neurophysiol 2015;32:324–30. doi:10.1097/WNP.0000000000000144.

Wiebe S. Review of The Johns Hopkins atlas of digital EEG: An interactive training guide. Can J Neurol Sci / Le J Can Des Sci Neurol 2008;35:269. doi:10.1001/archneurol.2007.44.

Wilson SB, Scheuer ML, Plummer C, Young B, Pacia S. Seizure detection: correlation of human experts. Clin Neurophysiol 2003;114:2156–64. doi:https://doi.org/10.1016/S1388-2457(03)00212-8.

Yamada T, Meng E. Practical guide for clinical neurophysiologic testing: EEG. Philadelphia, Pennsylvania, USA: Lippincott Williams & Wilkins; 2017. doi:10.1111%2Fj.1468-1331.2009.02936.x.

Ziyabari S, Shah V, Golmohammadi M, Obeid I, Picone J. Objective evaluation metrics for automatic classification of EEG events. J Neural Eng 2017:17. doi:https://www.isip.piconepress.com/publications/unpublished/journals/2018/jne/metrics.