Received: 25 July 2022 | Revised: 25 October 2022 | Accepted: 25 October 2022

DOI: 10.1111/epi.17446

Epilepsia®

CRITICAL REVIEW

Reliability of visual review of intracranial electroencephalogram in identifying the seizure onset zone: A systematic review and implications for the accuracy of automated methods

James Flanary[1] | Sam Daly[2] | Caitlin Bakker[3] | Alexander B. Herman[4] | Michael C. Park[5] | Robert McGovern[5] | Thaddeus Walczak[6] | Thomas Henry[6] | Theoden I. Netoff[7] | David P. Darrow[5,8]

[1]Department of Surgery, Walter Reed National Military Medical Center, Bethesda, Maryland, USA

[2]Department of Neurosurgery, Baylor Scott and White Health, Temple, Texas, USA

[3]Dr John Archer Library, University of Regina, Regina, Saskatchewan, Canada

[4]Department of Psychiatry, University of Minnesota, Minneapolis, Minnesota, USA

[5]Department of Neurosurgery, University of Minnesota, Minneapolis, Minnesota, USA

[6]Department of Neurology, University of Minnesota, Minneapolis, Minnesota, USA

[7]Department of Biomedical Engineering, University of Minnesota, Minneapolis, Minnesota, USA

[8]Department of Neurosurgery, Hennepin County Medical Center, Minneapolis, Minnesota, USA

Correspondence
David P. Darrow, Department of Neurosurgery, University of Minnesota, Minneapolis, MN, USA.
Email: darro015@umn.edu

Abstract

Visual review of intracranial electroencephalography (iEEG) is often an essential component for defining the zone of resection for epilepsy surgery. Unsupervised approaches using machine and deep learning are being employed to identify seizure onset zones (SOZs). This prompts a more comprehensive understanding of the reliability of visual review as a reference standard. We sought to summarize existing evidence on the reliability of visual review of iEEG in defining the SOZ for patients undergoing surgical workup and understand its implications for algorithm accuracy for SOZ prediction. We performed a systematic literature review on the reliability of determining the SOZ by visual inspection of iEEG in accordance with best practices. Searches included MEDLINE, Embase, Cochrane Library, and Web of Science on May 8, 2022. We included studies with a quantitative reliability assessment within or between observers. Risk of bias assessment was performed with QUADAS-2. A model was developed to estimate the effect of Cohen kappa on the maximum possible accuracy for any algorithm detecting the SOZ. Two thousand three hundred thirty-eight articles were identified and evaluated, of which one met inclusion criteria. This study assessed reliability between two reviewers for 10 patients with temporal lobe epilepsy and found a kappa of .80. These limited data were used to model the maximum accuracy of automated methods. For a hypothetical algorithm that is 100% accurate to the ground truth, the maximum accuracy modeled with a Cohen kappa of .8 ranged from .60 to .85 (*F*-2). The reliability of reviewing iEEG to localize the SOZ has been evaluated only in a small sample of patients with methodologic limitations. The ability of any algorithm to estimate the SOZ is notably limited by the reliability of iEEG interpretation. We acknowledge practical limitations of rigorous reliability

analysis, and we propose design characteristics and study questions to further investigate reliability.

KEYWORDS

electrocorticography, intracranial electroencephalography, reliability, seizure onset zone, stereoencephalography

1 | INTRODUCTION

Medically refractory epilepsy, when seizures are not adequately controlled through pharmacological treatment, represents approximately one third of patients seeking care for seizure treatment.[1] Certain types of medically refractory epilepsy have been shown to be amenable to surgical treatment, especially surgical resection when a seizure focus or seizure onset zone (SOZ) has been identified. The SOZ, or epileptogenic focus, is defined as the area of the brain in which seizures are generated or, pragmatically, as the leading nodes of the seizure network.[2] Clinically distinguishing this zone from normal tissue is a complex process, based on a concordance of findings from multiple investigations.[2] Although noninvasive localization methods using scalp electroencephalography (EEG) or magnetoencephalography have gained in popularity and sophistication,[3,4] they have not yet shown superiority to intracranial EEG (iEEG), also known as electrocorticography (ECoG), which has evolved to include both subdural cortical grids or strips and stereoencephalography.[5,6] iEEG is used to (1) determine whether there is an SOZ, (2) localize the SOZ, and (3) evaluate the proximity of the SOZ to eloquent structures that would carry significant morbidity if surgically resected or damaged. Identifying the SOZ can be difficult, and the best operative decision is not always clear.[7] The challenge of effective SOZ localization is highlighted by the substantial number of patients who have persistent seizures after resection; seizure freedom rates approach two thirds for temporal lobe epilepsy patients and one half for neocortical epilepsy patients at 2–5 years.[8] Multiple studies have attempted to describe markers, surrogates, or a "fingerprint" of the SOZ, but for each patient, its ultimate boundaries are subject to individual interpretation.[7,9] Interpretation of iEEG by an epileptologist remains a critical factor in the final decision regarding zone of resection.

Efforts have been made to algorithmically automate SOZ identification by comparing quantitative metrics against iEEG interpretation or postoperative seizure freedom following resection as reference standards.[10–16] Visual iEEG review is treated as a gold standard and is referenced as such.[10,13] An important limitation is that the sensitivity and specificity of unsupervised algorithmic studies are difficult to interpret when the reliability and validity of the comparison standard of visual iEEG review are not known. Comparing automated SOZ detection with postoperative outcomes has generally required a retrospective approach or the assumption that surgeries carry minimal error or uncertainty, although some have tried to incorporate the differences between algorithmic recommendations and surgeries through the probability of recurrence.[15–18] Seizure recurrence may be due to incomplete resection, secondary foci, or development of a new focus over time.[19] Quantifying the uncertainty along the workflow of surgical epilepsy will be critical to improving the long-term outcomes of those suffering from chronic epilepsy. A major gap in knowledge remains the uncertainty in the visual interpretation of iEEG to estimate a SOZ. As a gold standard, the final interpretation of iEEG from the preictal, ictal, and interictal periods to define the SOZ remains subject to human interpretation, with its corresponding biases, including surgical sampling bias and training bias.[7,20,21] Ideally, to serve as a gold standard, clinical review of iEEG by epileptologists would provide a highly valid and reliable diagnostic process for defining the SOZ. Whereas an improved mechanistic understanding of epilepsy may provide insight into the sensitivity and specificity of visual review of iEEG, measures of inter- and intrarater reliability reflect directly on the validity of a gold standard and are readily measurable.

Key Points

- Visual review of iEEG is an essential component in determining surgical boundaries during workup for nonlesional refractory epilepsy
- Insufficient data could be identified to support high agreement in visual review of iEEG
- SOZ detection algorithms play an increasingly important role, but visual interpretation of iEEG may limit the apparent maximum accuracy
- A study to investigate the reliability of SOZ localization can be performed and should include several considerations for study design

Inter- and intrarater reliability provide insight into the uncertainty inherent to a reference standard that relies on human interpretation. Several studies have explored the interrater reliability of scalp EEG in seizure localization.[22–27] One of these studies found that between the diagnostic categories of normal, ictal, and nonictal abnormalities, the probability of disagreement between a randomly selected pair of readers is approximately 23%, and there is at least a 11.5% probability of one reader being wrong about the abnormality.[23] However, differences in recording techniques and application mean that the reliability of scalp EEG is not analogous to iEEG. Scalp EEG is subject to volume conduction effects and obscuration by extracranial sources of artifacts that are minimized with iEEG.[28] Thus, variation in the reliability of scalp EEG localization of ictal discharges may be due to reviewer performance in interpretation domains that do not exist in iEEG. iEEG also involves a much greater number of electrodes focused on a narrow target, has a more focused goal in terms of operative planning, and can detect seizures not present on scalp EEG.[7,20] The arrangement of electrodes in iEEG in turn presents sampling bias, where electrodes are placed based on the best estimates of scalp EEG and imaging methods with a varying number of negative controls.[29,30]

Given the important implications of iEEG review and interpretation for clinical use, for surgical planning, and for understanding the limitations of automated algorithmic detection of SOZ, we sought to evaluate the current level of evidence of interrater reliability in localizing the SOZ from iEEG by systematically reviewing the literature. In addition, we planned to relate measures of reliability (e.g., Cohen kappa) to the maximum achievable accuracy of an ideal hypothetical algorithm through the use of a computational model.

2 | MATERIALS AND METHODS

A comprehensive literature search was performed across MEDLINE via PubMed, Embase via Ovid, the Cochrane Library, and Web of Science Core Collection to identify articles that evaluated reliability of epileptogenic focus localization using iEEG. Unpublished material and conference proceedings identified in these databases were included. In accordance with best practices, a combination of natural language searching and controlled vocabulary was used.[31] Search terms included criteria for seizures, iEEG, and reliability. Full search criteria are displayed in Table 1. Searches are up to date as of May 8, 2022. To ensure that no potentially relevant article was overlooked, the reference lists of all articles included for data extraction were searched to obtain additional articles. When articles appeared to refer to reliability analysis, but results were not reported, the authors of published studies were contacted for additional findings; this occurred on one occasion, and no additional data were received. This review was reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis) statement.[32] The protocol for this study was registered in Open Science Framework and is publicly available.[33]

Studies were included if they reported a quantitative assessment of reliability between at least two observers for SOZ localization with iEEG. Any method of assessing reliability was included as long as it pertained to SOZ. Reliability studies most commonly report a kappa value, a statistical measure of reliability in which the agreement and expected agreement by chance are considered. Measures of reliability are often between two reviewers and with nominal data; however, it can be applied for multiple reviewers and with quantitative data as well. Studies were excluded if they only assessed the interrater reliability of markers of interictal activity, such as epileptiform discharges, spikes, or high-frequency oscillations. Studies were also excluded if they evaluated the reliability of computer algorithms at detecting the SOZ, unless they evaluated reliability of human reviewers as a part of the process. There were no exclusion criteria based on sample size of patients, type of epilepsy, age of the patients, type or configuration of electrodes (i.e., depth vs. subdural electrodes), EEG sampling rates, duration of EEG data sampled, sample size of EEG segments, data processing techniques, whether reviewers were blinded to clinical data, statistical method of reliability analysis, publication year, or type of publication (i.e., article, conference proceeding, etc.). Non-English and nonhuman studies were excluded.

DistillerSR software was used for article screening and data extraction for all included articles. At each stage of the review and data extraction process, articles were reviewed independently by two authors (S.D. and J.F.) using inclusion and exclusion criteria as defined in the preceding paragraph. For title and abstract screening, disagreements were resolved by consensus, or advancement to the next level when consensus was not reached. For full text review, disagreements were resolved by consensus, and if needed, a third author served as tie-breaker (D.P.D.). During full-text screening, reasons for exclusion were recorded and are reported in Figure 1.

Piloted forms were used for both levels of article screening and for data extraction. Article screening forms were piloted by three authors (S.D., J.F., and D.P.D.). Data extraction forms were piloted by S.D. and J.F. and reviewed by S.D., J.F., and D.P.D. Data were extracted independently by S.D. and J.F. Extracted data

TABLE 1 Specific search terms used for each of the four databases searched

Database	Search date	Reliability terms		Seizure terms		ECoG terms
MEDLINE	May 8, 2022	(reliability[all fields] OR reliab*[all fields] OR "Reproducibility of results"[MeSH terms] OR "observer variation"[MeSH terms] OR "variability"[all fields] OR variab*[all fields] OR variation[all fields])	AND	("seizures"[MeSH terms] OR "seizures"[all fields] OR "seizure"[all fields] OR "epilepsy"[MeSH terms] OR "epilepsy"[all fields])	AND	("electrocorticography"[MeSH terms] OR "electrocorticography"[all fields] OR "ecog"[all fields] OR "intracranial EEG"[all fields] OR "iEEG"[all fields] OR "stereotactic EEG"[all fields] OR "sEEG"[all fields])
Embase	May 8, 2022	(reliability or reliab* or variab* or variation).mp. or (exp intrarater reliability/ or exp interrater reliability/ or exp test retest reliability/ or exp reliability/)	AND	(exp seizure/ or exp epilepsy/ or seizure.mp. or epilep*.mp.)	AND	(electrocorticography.mp. or exp electrocorticography/ or intracranial EEG.mp. or iEEG.mp. or stereotactic EEG.mp. or sEEG.mp.)
Cochrane Library	May 8, 2022	([reliability or reliab* or variab* or variation].mp. or [exp intrarater reliability/ or exp interrater reliability/ or exp test retest reliability/ or exp reliability/ or exp "observer variation"/ or exp "reproducibility of results"/])	AND	(exp seizure/ or exp epilepsy/ or seizure.mp. or epilep*.mp.)	AND	(electrocorticography.mp. or exp electrocorticography/ or intracranial EEG.mp. or iEEG.mp. or stereotactic EEG.mp. or sEEG.mp.)
Web of Science	May 8, 2022	ALL = (reliability* OR reproducibility of results OR observer variation)	AND	TS = (seizure* OR epilep*)	AND	([TS = (electrocorticograph* or intracranial EEG or iEEG or stereotactic EEG or sEEG)] AND DOCUMENT TYPES:[article])

Note: Each column was combined with "AND" in the search box.Abbreviation: ECoG, electrocorticography.

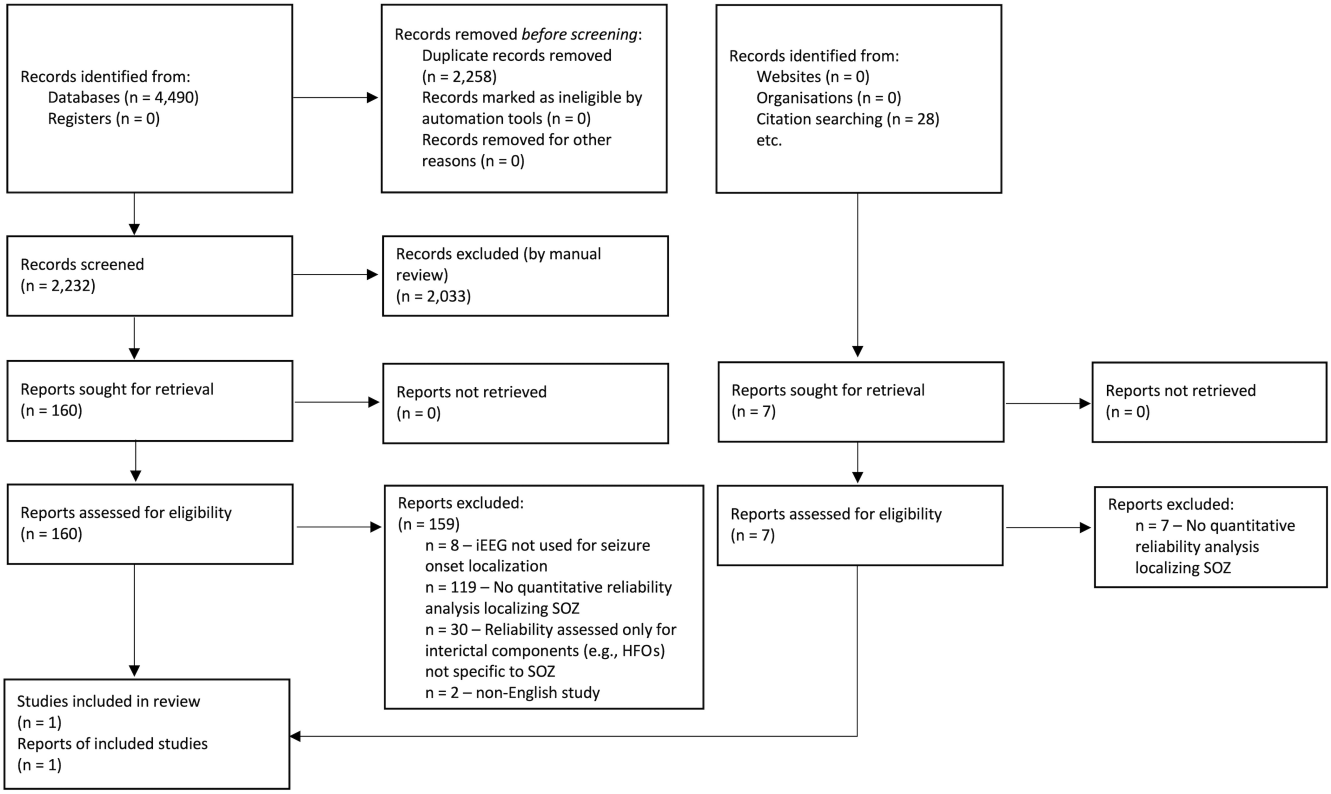

FIGURE 1 Diagram outlining included articles at each stage of the screening process, including reasons for exclusion at the full text screening stage. There were 21 articles identified through citation searching that were excluded via title and abstract screening. HFO, high-frequency oscillation; iEEG, intracranial electroencephalography; SOZ, seizure onset zone

included patient characteristics (number of patients, age, sex, type of epilepsy, zone of surgical resection, and clinical outcomes), all available iEEG data (including sampling method, location, and number of ECoG recordings), duration of recordings, number of seizures or segments, number of reviewers and training of reviewers, clinical information available to reviewers (i.e., blinding), method(s) of evaluating reliability, and all relevant reliability data. Study funding source, conflicts of interest, and any reported strengths and limitations were also recorded. Risk of bias was assessed for individual studies using QUADAS-2, a tool for quality assessment in accuracy studies for purposes of systematic review.[34] Bias was assessed independently by S.D. and J.F., and disagreements were resolved by consensus and with D.P.D. as tie-breaker when needed.

We anticipated significant heterogeneity of data with regard to measures of localization, duration, and context of iEEG data, and whether iEEG readers were blinded to clinical data. Not enough studies were identified to be able to consider a meta-analysis. As such, no sensitivity analysis, statistical assessment of publication bias, or assessment of heterogeneity was conducted. All data were tabulated as reported in the study without modification.

2.1 | Differences between protocol and review

The initial plan was to exclude studies with high risk of bias. Unfortunately, the only study had elements with high risk of bias. To address bias, we presented any relevant study characteristics that introduce bias with the results. The paucity of eligible studies also prompted revision of search criteria. It was decided to add an additional database, Web of Science Core Collection, to ensure identification of all relevant studies.

2.2 | Estimation of *F*-beta as a function of kappa

In addition to characterizing the existing literature on estimates of rater reliability, a main goal of this study was to relate the impact of imperfect reliability on the maximum achievable accuracy of an algorithm attempting to determine the SOZ. Even if a new algorithm perfectly predicts the ground truth SOZ, a gold standard that has a low kappa will limit maximum accuracy of the algorithm to standard. One metric that is commonly used for accuracy is the *F* score (also known as *F*-beta or *F* measure),

which is derived from the precision and sensitivity of a test. When beta = 1, precision and sensitivity are weighted equally. A beta of 2 (*F*-2) is commonly used when sensitivity is considered more important than precision, such as when it is important to not miss an area of the SOZ.[35] To understand how interrater reliability (Cohen kappa) affects values of accuracy, such as *F*-beta, we developed a computational model representing SOZ detection true positive and false positive rates and the expected *F*-beta and kappa rates.

F-beta is defined as:

$$F_\beta = \frac{\left(1+\beta^2\right)\times \mathrm{TP}}{\left(1+\beta^2\right)\times \mathrm{TP}+\beta^2\times \mathrm{FN}+\mathrm{FP}}$$

Where TP is the true positive, the number of electrodes in the SOZ that were detected in the SOZ. FN represents the number of false negative electrodes, that is, electrodes that were in the SOZ but were not identified. FP is the false positive, the number of electrodes outside the SOZ identified as inside the SOZ. The parameter β can be used to balance the weight of TPs to FNs. If $\beta = 1$, they are equally weighted. If $\beta = 2$, more weight is put on identifying the SOZ and accepting FPs than missing electrodes in the focus. For this study, we used $\beta = 2$, which we have also previously used.[36]

Interrater reliability, Cohen kappa, and measures of accuracy are dependent on the probability of TP's, detection of electrodes within the SOZ (pTP) and the probability of TN's, detection of electrodes outside the SOZ (pTN). These probabilities depend on the true proportion of electrodes within the SOZ. For simplicity we assume there are a total of 100 channels/contacts of iEEG, and 10 of the contacts are within the SOZ. We model all possible combinations of accuracies of the raters of pTP and pTN over a range from .6 to 1.0 and estimate the average value of kappa, and the average value of F-beta.

Modeling: N, # of electrodes; S, # electrodes in SOZ; $O = N - S$, # electrodes outside the SOZ; $pE = \left(\frac{S}{N}\right)^2$, # likelihood of agreement at random; $A_S = \mathrm{pTP}^2 S + (1-\mathrm{pTN})^2 O$, expected number of electrodes in SOZ that are agreed upon by two reviewers; $A_O = pTN^2 O + (1-pTN)^2 S$, expected number of electrodes outside the SOZ that are agreed upon by the two reviewers. The expected kappa could be calculated as:

$$E(kappa) = \frac{\frac{A_S + A_O}{N} - pE}{1 - pE},$$

and the expected F_β is calculated as:

$$E\left(F_\beta\right) = \frac{\left(1-\beta^2\right)(\mathrm{pTP}\times S)}{\left(1+beta^2\right)(\mathrm{pTP}\times S)+\beta^2(S-\mathrm{pTP}\times S)+(O-\mathrm{pTN}\times O)}.$$

3 | RESULTS

3.1 | Literature review

Titles and abstracts for 2338 articles were screened for data on iEEG reliability from four databases and the reference lists of included articles. One article ultimately met inclusion criteria (Figure 1). Characteristics and outcomes are reported in Table 2. The study had high risk of bias in patient selection and unclear risk of bias in application of the reference standard (Table 3).

The only article that met inclusion criteria for this study assessed reliability between two reviewers who interpreted 67 seizures from 10 patients with temporal lobe epilepsy who underwent iEEG analysis and subsequent surgical resection.[37] Reviewers marked the included channels of the SOZ for patients already determined to have an ictal onset that corresponded to a defined anatomic region. In this context, they observed a rater agreement of .97 and Cohen kappa of .80. This analysis was done as part of a study evaluating the effectiveness of an algorithm predicting the seizure focus. The specifics of what constituted reliable localization were not described.

There were several articles (detailed in Table S1) identified that did not meet our inclusion criteria but are still relevant in the discussion of reliability in iEEG interpretation. Haut et al. assessed reliability of seizure localization across 11 patients, almost exclusively with temporal lobe epilepsy.[38] It assessed reliability of the multidisciplinary conference decision made jointly by a center, using all standard modalities. These included extracranial EEG, clinical findings, magnetic resonance imaging (MRI), positron emission tomography (PET), and neuropsychologic testing in addition to iEEG. In this context, they reported "excellent" reliability in localization, with intraclass correlation coefficient (ICC) of .79.

Park et al. examined agreement between side of onset, focal versus regional onset, and pattern of seizure onset.[39] This study reported agreement in focal versus regional determination of 96%. Perucca et al. examined reliability of seizure onset time and morphology between two reviewers, blind to clinical data, of 33 patients with a focal structural lesion.[40] Observed kappa was .68 (.53–.83).

Two studies assessed reliability of "seizure vs. not seizure," both of which assessed the performance of implantable long-term monitoring devices and thus were limited by four to eight electrodes.[41,42] In Quigg et al., five reviewer pairs interpreted 7221 segments from 128 patients from an implanted system that consisted of four bipolar channels. The Cohen kappa for determining whether the segment contained a seizure event was .57 (Table 2).[41] Osorio et al. focused on assessing the algorithm of a long-term implantable monitor, which included up to eight channels

TABLE 2 Known patient demographics, study design methods (including iEEG recording methods), and relevant outcomes pertaining to iEEG interrater reliability

Study	Vila-Vidal et al., 2020
Patient demographics	
n	10 (9 underwent intracranial monitoring)
Age, mean years	36.2
Male, %	50
Type of disease	Nonlesional temporal lobe epilepsy. Inclusion was based on the following criteria: 1. that the seizure focus had been identified by the epileptologists 2. that ictal onset was confined to a reduced number of contacts corresponding to an anatomical region
Zone of resection (*n*)	TATL (5) SAH (1) RF-TC (3) NO (1)
Clinical outcomes, *n*	6 patients: Engel I: 5 Engel III: 1
Study design	
iEEG data: sampling method, location, and number of recordings	906 total recording electrodes; 5–21 intracerebral multiple contact microelectrodes; between 56–126 contacts per patient; 500-Hz sampling rate (except one patient with 250 Hz); a broadband pass filter (1165 Hz) and FIR notch filter (50 Hz) were used
Duration of segment	Entire seizure duration (not specified)
Number of seizures/segments reviewed, total	67 seizures
Reviewers, *n*	2
Training of reviewers	Unclear
Blinding	Unclear
Method(s) of evaluating reliability	Reviewers marked seizure onset zone-included channels; 3–14 were marked; reliability was assessed with Cohen kappa
Results	
Interrater reliability of iEEG	For seizure onset zone localization: • Percent interrater agreement: 97% • Cohen kappa: .80

Abbreviations: FIR, finite impulse response; iEEG, intracranial electroencephalography; NO, not operated; RF-TC, radiofrequency thermocoagulation; SAH, selective amygdalohippocampectomy; TATL, tailored anterior temporal lobectomy.

from depth and grid electrodes. As part of the analysis of their algorithm, they separately reported agreement of a subset of iEEG segments in terms of seizures, epileptiform discharges, and physiologic activity/artifact.[42]

3.2 | Computational modeling of kappa

Both the expected kappa and *F*-beta values can be modeled as a function of the probability of TP electrodes within the SOZ (pTP) and probability of selecting TN electrodes outside the SOZ (pTN; Figure 2). In an example in which there are 10 electrodes within the SOZ and 90 outside, a kappa of .8 can be observed across values of pTP without significant change, although this varies substantially based on small values of pTN. These proportions are more equal in contributing to expected kappa. This shows that in an example where kappa = .8, the expected beta can vary from .6 to .85 depending on the pTP rate. Therefore, a perfect algorithm that achieves 100% accuracy relative to ground truth will have a maximum *F*-beta of .85 when compared to a gold standard with a kappa of .8.

4 | DISCUSSION

We systematically reviewed the literature to identify studies examining the reliability of iEEG to determine the SOZ. Despite screening more than 2300 papers, we identified only one study that assessed the reliability of seizure focus localization, reporting a Cohen kappa of .8. No study was identified whose primary purpose was analyzing reliability in localization of the SOZ. Our computational model indicates that a Cohen kappa of .8 would result in a range of accuracies from .6 to .85 (as measured by *F*-2), assuming the algorithm has 100% accuracy in detecting the ground truth SOZ. Interrater uncertainty in this gold standard confers limitations in the maximal achievable accuracy by any algorithmic approach.

Although a Cohen kappa of .8 is similar to kappa scores reported for scalp EEG, there are several limitations. Most significant is the study population and size, which were limited to 10 participants, known to have ictal onset within a defined anatomic region, with temporal lobe epilepsy. It was not clear whether blinding was performed from adjunct clinical data, such as MRI, PET, or multidisciplinary conference data. For extratemporal lobe epilepsy, discharges are highly variable and poorly understood, and seizure-free outcomes are lower.[8,43,44] For these patients, iEEG may play a critical role in SOZ identification; however, interrater reliability remains unknown.

As technological improvements fuel safer surgery, increased contact density, less invasive approaches, and new

TABLE 3 Risk of bias according to the QUADAS-2 tool

Study	Risk of bias			Applicability concerns	
	Patient selection	Reference standard	Flow and timing	Patient selection	Reference standard
Vila-Vidal et al. (2020)	High	Unclear	Low	Low	Unclear

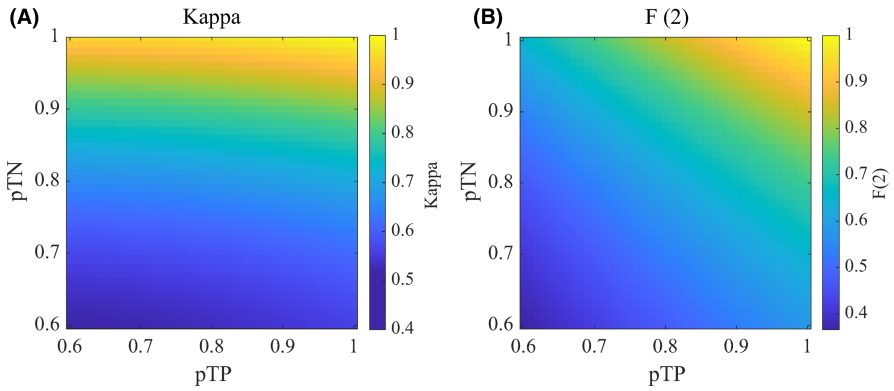

FIGURE 2 The expected kappa (A), and expected beta (B) for a hypothetical study assessing reliability of a seizure foci prediction algorithm that uses epileptologist-defined seizure onset zone as the gold standard. Both are expressed as a function of the probability of true positive electrodes within the seizure onset zone (pTP) and probability of selecting true negative electrodes outside the seizure onset zone (pTN)

neuromodulation platforms, effective algorithm development and implementation will be critical. Ideally, a gold standard should provide an accurate, unbiased, and reliable result. Characterizing the reliability of iEEG to identifying SOZs is paramount in interpreting the utility of algorithms. By modeling the basic parameters linking interrater reliability and a measure of accuracy, we demonstrate how uncertainty in reliability can directly contribute to uncertainty in accuracy of algorithms, even in a disproportionate way where a kappa of .8 can result in an *F*-2 of .6.

Our interdisciplinary group recognizes the challenges in designing and implementing a reliability study for iEEG, including sampling, training, and institutional biases, among others. A proposed study outline is presented in PICO format in Table 4. The ideal study would be blinded to avoid sampling bias, include both temporal lobe and extratemporal epilepsy, and involve many institutions to overcome training bias and institution bias. For appropriate statistical analysis, there should be an adequate quantity of patients and epileptologists. Sample size estimates according to proportion of positive ratings and estimated kappa are described elsewhere. To reject a null kappa of at least .60, a minimum of 56 patients is required, although the number can increase substantially depending on alternative criteria.[45] Both kappa and ICC can be applied (i.e., kappa is used for categorical variables, whereas ICC is used for continuous quantitative variables); each has limitations based on application and can be chosen according to the study context.[46]

Stratification of cases by indications for iEEG may improve interpretation of reliability analyses, in part by elucidating the effects of these indications on intracranial electrode selection and success in placing electrodes near the actual SOZ.[47] For example, patients with single epileptogenic lesions but poorly localized scalp EEG ictal onsets may have electrodes implanted near their particular lesion as a presumptive SOZ. By contrast, patients with no lesions on imaging and with poorly localized scalp EEG ictal onsets may not have intracranial electrodes placed in the SOZ. In cases such as this, sites of early propagation of ictal discharges may have the earliest detectable seizure discharges in a more widely distributed topography; presumably such situations would degrade consensus among reviewers. Thus, in addition to analyzing effects of temporal versus extratemporal SOZs on reliability of qualitative iEEG SOZ localization, the indications for iEEG should be considered. It would also be worthwhile to include seizure-free surgical outcomes as an additional reference standard.

There is significant precedent for improving our understanding of a reference standard through means such as composite reference standards, alternative tests, or panel consensus.[48] The development of consensus guidelines for EEG interpretation of scalp EEG, for example, was shown to improve interrater reliability.[26] This represents one of many options that could be applied to iEEG. It has been suggested that given the distribution of surgical epilepsy

Patients	Consecutive involvement of epilepsy patients undergoing evaluation for resective surgery. No restriction on age or type of epilepsy. Power analysis to allow for separate analysis of temporal and extratemporal epilepsy.
Intervention	Subdural or depth electrodes in any configuration. EEG readers provided access to video EEG. EEG readers blinded to imaging and multidisciplinary conference recommendations.
Comparison	Not applicable.
Outcome	Reliability analyses: • Initial anatomic region of seizure onset. • Number of electrodes in the seizure onset zone. • Anatomical region involved in seizure propagation outside the initial region of onset. • Interval between seizure onset and propagation outside of initial region. • Time of electrographic seizure onset. • Time of clinical seizure onset. • Interval between onset of electrographic seizure and clinical seizure. • Time of electrographic seizure offset.

Abbreviation: EEG, electroencephalography.

TABLE 4 Proposed PICO diagram for a multicenter reliability study including multiple measurable and reproducible criteria

patients across many centers, developing common methods and means of reporting, such as via the National Institutions of Health-sponsored Common Data Elements or other techniques, may facilitate cooperation or meta-analyses.[49] A starting point would be a multicenter reliability study following practices described above.

4.1 | Limitations

The primary purpose of this review is to highlight the current level of evidence supporting visual review as the gold standard for SOZ identification in iEEG through the lens of reliability. High reliability between reviewers remains a critical component of any diagnostic test, especially when surgical recommendations are made. High reliability should also be balanced by improved surgical outcomes, although multiple other factors may play a role in determining postoperative seizure recurrence that are not discernible from using iEEG to identify the SOZ. Although our search was exhaustive, reliability has been a secondary concern for studies that analyze it, so some discreetly reported results may be missing from this systematic review. As such, few studies were found, which hampered further analysis and our final interpretations.

5 | CONCLUSIONS

Although iEEG remains the standard in identification of the SOZ for many patients, remarkably few studies have assessed the reliability of its use in this context. The only study identified was not focused specifically on this question and has substantial limitations in generalizability for many patients undergoing iEEG. As algorithms are increasingly used in attempts to predict localization of the SOZ, the limitations in knowledge of reviewer reliability should be considered. We believe this presents an important area for further research.

ACKNOWLEDGMENTS

None.

CONFLICT OF INTEREST

None of the authors has any conflict of interest to disclose. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

ORCID

James Flanary https://orcid.org/0000-0001-8206-0469
Sam Daly https://orcid.org/0000-0002-3652-491X
Caitlin Bakker https://orcid.org/0000-0003-4154-8382
Robert McGovern https://orcid.org/0000-0003-0752-1899
David P. Darrow https://orcid.org/0000-0001-9335-0584

REFERENCES

1. Laxer KD, Trinka E, Hirsch LJ, Cendes F, Langfitt J, Delanty N, et al. The consequences of refractory epilepsy and its treatment. Epilepsy Behav. 2014;37:59–70.

2. Siegel AM. Presurgical evaluation and surgical treatment of medically refractory epilepsy. Neurosurg Rev. 2004;27(1):1–18; discussion 19–21.
3. Knowlton RC. The role of FDG-PET, ictal SPECT, and MEG in the epilepsy surgery evaluation. Epilepsy Behav. 2006;8(1):91–101.
4. Englot DJ, Nagarajan SS, Imber BS, Raygor KP, Honma SM, Mizuiri D, et al. Epileptogenic zone localization using magnetoencephalography predicts seizure freedom in epilepsy surgery. Epilepsia. 2015;56(6):949–58.
5. Aubert S, Wendling F, Regis J, McGonigal A, Figarella-Branger D, Peragut J-C, et al. Local and remote epileptogenicity in focal cortical dysplasias and neurodevelopmental tumours. Brain. 2009;132(Pt 11):3072–86.
6. Kahane P, Landré E, Minotti L, Francione S, Ryvlin P. The Bancaud and Talairach view on the epileptogenic zone: a working hypothesis. Epileptic Disord. 2006;8(Suppl 2):S16–26.
7. Rosenow F, Lüders H. Presurgical evaluation of epilepsy. Brain. 2001;124(Pt 9):1683–700.
8. Engel J Jr, Wiebe S, French J, Sperling M, Williamson P, Spencer D, et al. Practice parameter: temporal lobe and localized neocortical resections for epilepsy: report of the Quality Standards Subcommittee of the American Academy of Neurology, in association with the American Epilepsy Society and the American Association of Neurological Surgeons. Neurology. 2003;60(4):538–47.
9. Grinenko O, Li J, Mosher JC, Wang IZ, Bulacio JC, Gonzalez-Martinez J, et al. A fingerprint of the epileptogenic zone in human epilepsies. Brain. 2018;141(1):117–31.
10. Liu S, Sha Z, Sencer A, Aydoseli A, Bebek N, Abosch A, et al. Exploring the time-frequency content of high frequency oscillations for automated identification of seizure onset zone in epilepsy. J Neural Eng. 2016;13(2):026026.
11. Burrello A, Schindler K, Benini L, Rahimi A. Hyperdimensional computing with local binary patterns: one-shot learning of seizure onset and identification of ictogenic brain regions using short-time iEEG recordings. IEEE Trans Biomed Eng. 2020;67(2):601–13.
12. Varatharajah Y, Berry B, Cimbalnik J, Kremen V, Van Gompel J, Stead M, et al. Integrating artificial intelligence with real-time intracranial EEG monitoring to automate interictal identification of seizure onset zones in focal epilepsy. J Neural Eng. 2018;15(4):046035.
13. Murin Y, Kim J, Parvizi J, Goldsmith A. SozRank: a new approach for localizing the epileptic seizure onset zone. PLoS Comput Biol. 2018;14(1):e1005953.
14. Elahian B, Yeasin M, Mudigoudar B, Wheless JW, Babajani-Feremi A. Identifying seizure onset zone from electrocorticographic recordings: a machine learning approach based on phase locking value. Seizure. 2017;51:35–42.
15. Park E-H, Madsen JR. Granger causality analysis of interictal iEEG predicts seizure focus and ultimate resection. Neurosurgery. 2017;82(1):99–109. https://doi.org/10.1093/neuros/nyx195
16. Park S-C, Chung CK. Postoperative seizure outcome-guided machine learning for interictal electrocorticography in neocortical epilepsy. J Neurophysiol. 2018;119(6):2265–75.
17. Dian JA, Colic S, Chinvarun Y, Carlen PL, Bardakjian BL. Identification of brain regions of interest for epilepsy surgery planning using support vector machines. Annu Int Conf IEEE Eng Med Biol Soc. 2015;2015:6590–3.
18. Abbasi B, Goldenholz DM. Machine learning applications in epilepsy. Epilepsia. 2019;60(10):2037–47.
19. Jehi L, Yehia L, Peterson C, Niazi F, Busch R, Prayson R, et al. Preliminary report: late seizure recurrence years after epilepsy surgery may be associated with alterations in brain tissue transcriptome. Epilepsia Open. 2018;3(2):299–304.
20. Parvizi J, Kastner S. Promises and limitations of human intracranial electroencephalography. Nat Neurosci. 2018;21(4):474–83.
21. Jayakar P, Gotman J, Harvey AS, Palmini A, Tassi L, Schomer D, et al. Diagnostic utility of invasive EEG for epilepsy surgery: indications, modalities, and techniques. Epilepsia. 2016;57(11):1735–47.
22. Walczak TS, Radtke RA, Lewis DV. Accuracy and interobserver reliability of scalp ictal EEG. Neurology. 1992;42(12):2279–85.
23. Grant AC, Abdel-Baki SG, Weedon J, Arnedo V, Chari G, Koziorynska E, et al. EEG interpretation reliability and interpreter confidence: a large single-center study. Epilepsy Behav. 2014;32:102–7.
24. Halford JJ, Pressly WB, Benbadis SR, Tatum WO 4th, Turner RP, Arain A, et al. Web-based collection of expert opinion on routine scalp EEG: software development and interrater reliability. J Clin Neurophysiol. 2011;28(2):178–84.
25. Benbadis SR, LaFrance WC Jr, Papandonatos GD, Korabathina K, Lin K, Kraemer HC, et al. Interrater reliability of EEG-video monitoring. Neurology. 2009;73(11):843–6.
26. Azuma H, Hori S, Nakanishi M, Fujimoto S, Ichikawa N, Furukawa TA. An intervention to improve the interrater reliability of clinical EEG interpretations. Psychiatry Clin Neurosci. 2003;57(5):485–9.
27. Risinger MW, Engel J Jr, Van Ness PC, Henry TR, Crandall PH. Ictal localization of temporal lobe seizures with scalp/sphenoidal recordings. Neurology. 1989;39(10):1288–93.
28. Gloor P. Neuronal generators and the problem of localization in electroencephalography. J Clin Neurophysiol. 1985;2(4):327–54.
29. Kennedy BC, Katz J, Lepard J, Blount JP. Variation in pediatric stereoelectroencephalography practice among pediatric neurosurgeons in the United States: survey results. J Neurosurg Pediatr. 2021;28(2):1–9.
30. Mercier MR, Dubarry A-S, Tadel F, Avanzini P, Axmacher N, Cellier D, et al. Advances in human intracranial electroencephalography research, guidelines and good practices. Neuroimage. 2022;260:119438.
31. Churchill R, Lasserson T, Chandler J, Tovey D, Higgins J. Standards for the reporting of new Cochrane intervention reviews. In: Higgins JPT, Lasserson T, Chandler J, Tovey D, Churchill R, editors. Methodological expectations of Cochrane intervention reviews. London, UK: Cochrane; 2022.
32. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160.
33. Flanary J. Reliability of seizure localization in iEEG. 2020. https://osf.io/cw4m2/. Accessed 24 Jul 2022.
34. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36.
35. Van Rijsbergen CJ. Information retrieval. Oxford, UK: Butterworth-Heinemann; 1979.

36. Park Y, Luo L, Parhi KK, Netoff T. Seizure prediction with spectral power of EEG using cost-sensitive support vector machines. Epilepsia. 2011;52(10):1761–70.
37. Vila-Vidal M, Pérez Enríquez C, Principe A, Rocamora R, Deco G, Tauste CA. Low entropy map of brain oscillatory activity identifies spatially localized events: a new method for automated epilepsy focus prediction. Neuroimage. 2020;208::116410.
38. Haut SR, Berg AT, Shinnar S, Cohen HW, Bazil CW, Sperling MR, et al. Interrater reliability among epilepsy centers: multicenter study of epilepsy surgery. Epilepsia. 2002;43(11):1396–401.
39. Park YD, Murro AM, King DW, Gallagher BB, Smith JR, Yaghmai F. The significance of ictal depth EEG patterns in patients with temporal lobe epilepsy. Electroencephalogr Clin Neurophysiol. 1996;99(5):412–5.
40. Perucca P, Dubeau F, Gotman J. Intracranial electroencephalographic seizure-onset patterns: effect of underlying pathology. Brain. 2014;137(Pt 1):183–96.
41. Quigg M, Sun F, Fountain NB, Jobst BC, Wong VSS, Mirro E, et al. Interrater reliability in interpretation of electrocorticographic seizure detections of the responsive neurostimulator. Epilepsia. 2015;56(6):968–71.
42. Osorio I, Frei MG, Giftakis J, Peters T, Ingram J, Turnbull M, et al. Performance reassessment of a real-time seizure-detection algorithm on long ECoG series. Epilepsia. 2002;43(12):1522–35.
43. Englot DJ, Breshears JD, Sun PP, Chang EF, Auguste KI. Seizure outcomes after resective surgery for extra–temporal lobe epilepsy in pediatric patients. J Neurosurg Pediatr. 2013;12(2):126–33.
44. Haglund MM, Ojemann GA. Extratemporal resective surgery for epilepsy. Neurosurg Clin N Am. 1993;4(2):283–92.
45. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005;85(3):257–68.
46. Hernaez R. Reliability and agreement studies: a guide for clinical investigators. Gut. 2015;64(7):1018–27.
47. Henry TR, Ross DA, Schuh LA, Drury I. Indications and outcome of ictal recording with intracerebral and subdural electrodes in refractory complex partial seizures. J Clin Neurophysiol. 1999;16(5):426–38.
48. Reitsma JB, Rutjes AWS, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol. 2009;62(8):797–806.
49. Quigg M. The reliability of intraoperative electrocorticography in magnetic resonance imaging-negative temporal lobe epilepsy: spikes mark the spot. JAMA Neurol. 2014;71(6):681–2.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Flanary J, Daly S, Bakker C, Herman AB, Park MC, McGovern R, et al. Reliability of visual review of intracranial electroencephalogram in identifying the seizure onset zone: A systematic review and implications for the accuracy of automated methods. Epilepsia. 2023;64:6–16. https://doi.org/10.1111/epi.17446