Received: 15 August 2019 | Revised: 1 November 2019 | Accepted: 5 November 2019

DOI: 10.1111/epi.16397

FULL-LENGTH ORIGINAL RESEARCH

Epilepsia®

Prospective validation study of an epilepsy seizure risk system for outpatient evaluation

Sharon Chiang[1] | **Daniel M. Goldenholz**[2] | **Robert Moss**[3] | **Vikram R. Rao**[1] | **Zulfi Haneef**[4,5] | **William H. Theodore**[6] | **Jonathan K. Kleen**[1] | **Jay Gavvala**[4] | **Marina Vannucci**[7] | **John M. Stern**[8]

[1]Department of Neurology and Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, California

[2]Department of Neurology, Beth Israel Deaconess Medical Center, Boston, Massachusetts

[3]Seizure Tracker TM LLC, Annandale, Virginia

[4]Department of Neurology, Baylor College of Medicine, Houston, Texas

[5]Neurology Care Line, VA Medical Center, Houston, Texas

[6]Clinical Epilepsy Section, National Institute of Neurological Disorders and Stroke (NINDS), National Institutes of Health (NIH), Bethesda, Maryland

[7]Department of Statistics, Rice University, Houston, Texas

[8]Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California

Correspondence
Sharon Chiang, Department of Neurology and Weill Institute for Neurosciences, University of California, San Francisco, 505 Parnassus Avenue, San Francisco, CA 94143.
Email: Sharon.Chiang@ucsf.edu

Abstract

Objective: We conducted clinical testing of an automated Bayesian machine learning algorithm (Epilepsy Seizure Assessment Tool [EpiSAT]) for outpatient seizure risk assessment using seizure counting data, and validated performance against specialized epilepsy clinician experts.

Methods: We conducted a prospective longitudinal study of EpiSAT performance against 24 specialized clinician experts at three tertiary referral epilepsy centers in the United States. Accuracy, interrater reliability, and intra-rater reliability of EpiSAT for correctly identifying changes in seizure risk (improvements, worsening, or no change) were evaluated using 120 seizures from four synthetic seizure diaries (seizure risk known) and 120 seizures from four real seizure diaries (seizure risk unknown). The proportion of observed agreement between EpiSAT and clinicians was evaluated to assess compatibility of EpiSAT with clinical decision patterns by epilepsy experts.

Results: EpiSAT exhibited substantial observed agreement (75.4%) with clinicians for assessing seizure risk. The mean accuracy of epilepsy providers for correctly assessing seizure risk was 74.7%. EpiSAT accurately identified seizure risk in 87.5% of seizure diary entries, corresponding to a significant improvement of 17.4% ($P = .002$). Clinicians exhibited low-to-moderate interrater reliability for seizure risk assessment (Krippendorff's $\alpha = 0.46$) with good intrarater reliability across a 4- to 12-week evaluation period (Scott's $\pi = 0.89$).

Significance: These results validate the ability of EpiSAT to yield objective clinical recommendations on seizure risk which follow decision patterns similar to those from specialized epilepsy providers, but with improved accuracy and reproducibility. This algorithm may serve as a useful clinical decision support system for quantitative analysis of clinical seizure frequency in clinical epilepsy practice.

KEYWORDS

clinical decision support system, interrater reliability, intrarater reliability, seizure risk

1 | INTRODUCTION

Nearly one-third of referrals in neurology are for new-onset or breakthrough seizures.[1–3] Neurologists are often provided with the date of a breakthrough seizure or increase in seizure frequency, and asked to determine whether this reflects an actual change in the underlying propensity toward seizures warranting intervention change**.** However, interpreting changes in raw frequencies of patient-reported seizures involves a degree of uncertainty.[4] Crude estimates of seizure frequency over the past few months can be misleading; often, apparent changes in seizure frequency can be caused by the simple natural variability of epilepsy or variations in reporting accuracy.[5–8] If mistakenly interpreted as a change in the underlying propensity toward seizures, this may lead to unnecessary or potentially harmful treatment changes.

A decision support algorithm to improve quantitative interpretation of seizure frequencies and guide identification of periods of heightened seizure propensity is needed. Raw seizure frequencies alone are not sufficient for judging when patients are in a state of heightened seizure susceptibility: for example, is an increase from three to five seizures per month meaningful, is this natural variation, or is this possibly just a discrepancy or discounted omission in patient recording? Current clinical practice employed by neurologists and epileptologists involves a large degree of subjectivity in making such determinations. Intracranial electroencephalography (EEG) has confirmed the presence of distinct "proictal" states in patients with epilepsy, or brain states of increased seizure susceptibility,[9–12] and provides an exciting new potential method for objectively identifying brain states of heightened seizure susceptibility based on quantitative intracranial electrographic criteria. New evidence shows that not only electrographic but also clinical seizures are observed during specific phases of these high seizure susceptibility states.[10,13] This suggests that it may be possible to decode underlying changes in seizure susceptibility using noninvasive clinical seizure frequencies.

This study pursues the second stage of testing of a new quantitative tool for seizure risk evaluation (Epilepsy Seizure Assessment Tool [EpiSAT]), which allows automatic identification of times when changes in patient-reported clinical seizure frequency are at high probability for indicating real worsening or improvement in seizure susceptibility. Previous testing has demonstrated analytical validity using the SeizureTracker.com database and shown that compared to simulated constructs of physician decision-making, EpiSAT more accurately uses raw clinical seizure frequencies to decode changes in underlying seizure propensity states**.**[14] In this study, we evaluate the performance of EpiSAT compared to specialized human epilepsy experts using a prospective longitudinal study design at three National Association of Epilepsy Centers (NAECs; level 4 epilepsy centers) in the United States.

Key Points

- Epilepsy clinicians exhibited high intrarater reliability (Scott's $\pi = 0.89$) but low-to-moderate interrater reliability (Krippendorff's $\alpha = 0.45$) in using raw seizure frequencies to judge change improvements or worsening in underlying seizure risk.
- The EpiSAT algorithm exhibits substantial observed agreement (75.4%) in decision patterns with human epilepsy clinicians
- Use of EpiSAT may provide a useful tool for quantitative interpretation of seizure frequencies to identify periods of heightened seizure propensity in people with epilepsy

2 | METHODS

2.1 | Subjects

American Board of Psychiatry and Neurology (ABPN) epilepsy board-certified epileptologists, nurse practitioners, and clinical neurophysiology/epilepsy fellows in adult epilepsy practice were recruited via advertisement and word of mouth at the University of California San Francisco, Beth Israel Deaconess Medical Center and Baylor College of Medicine between January 2019 and May 2019. Informed consent was obtained from all providers. A principal investigator at each site (Z.H., D.M.G., V.R.R., S.C.) sent the study questionnaire to each provider. To enhance response rates, a second round of follow-up emails was sent to non-responders. Of the 31 attending physicians, four nurse practitioners, and eight fellows recruited, we had responses from a total of 24 epilepsy providers (77.4% response rate). For each respondent, the number of years experience, position, and whether the provider differentiated between seizure risk and random fluctuations in seizure frequency in everyday clinical practice was recorded. Written informed consent was obtained from each participant. This study was approved by the institutional review board of each institution.

2.2 | Seizure risk system

The EpiSAT system is a new seizure risk machine-learning algorithm for identifying and assessing changes in seizure

risk.[14] In brief, EpiSAT automatically decodes raw seizure frequencies into seizure susceptibility states, by modeling raw seizure frequency in any specified time unit as the observed manifestation of a time-varying hidden (unobserved) seizure risk state using a Bayesian point process (Figure S1). The unobserved seizure risk state takes ordinal values and is used to capture brain states of heightened seizure susceptibility based on the temporal patterns in raw seizure frequencies. Temporal dependencies are captured using a hidden Markov process, which models the temporal dependency between seizure frequencies as dependence in underlying factors producing epilepsy, including seizure threshold, epileptogenic abnormalities, and precipitating factors.[15] Finally, the model allows external clinical measurements to affect the probability that a patient will worsen/improve to a higher/lower seizure risk state. Bayesian techniques are used to update hidden transition probabilities based on observed data (Figure S2). This solution allows missed data to be incorporated in estimations based on the recognition of reproducible patterns underlying seizure activity, which allows for robustness to error or omission, with <50% error even with up to 70% missing data.[14] (For algorithm details, see Appendix S1.)

2.3 | Prospective testing format

ABPN epilepsy board-certified epileptologists and nurse practitioners were presented with two synthetic patient seizure diaries and two real patient seizure diaries, and epilepsy fellows were presented with four synthetic patient seizure diaries and four real patient seizure diaries, using a multipage web survey to distribute cases. A larger number of cases were presented to fellows to account for anticipated greater variability with fewer years of subspecialty experience. Each study questionnaire contained monthly

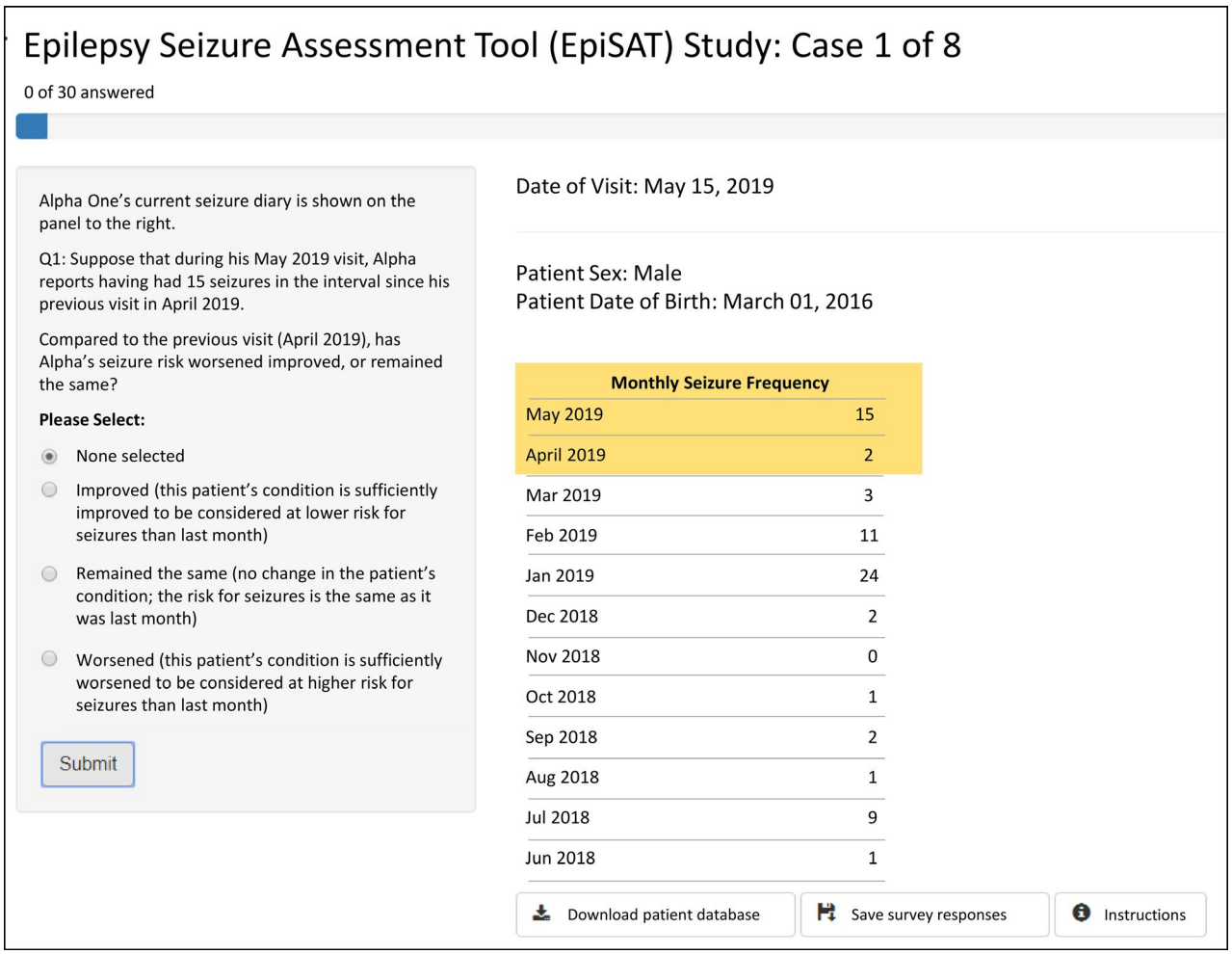

FIGURE 1 Sample page from seizure diary web survey completed by providers. Each page presented the provider with the patient's most recent seizure diary entry, as well as historical seizure diary entries and a database of similar patients. Seizure diary entries were revealed sequentially. For each entry, providers were asked to determine whether, compared to the patient's previous entry, the patient had worsened, improved, or remained the same in underlying seizure risk. The same set of cases was repeated 4 to 12 weeks following completion of the initial cases

seizure diary entries recorded in SeizureTracker.com by a person with epilepsy who consented to the research. For each patient case, clinicians were presented with the patient's most recent monthly seizure frequency, as well as monthly seizure frequencies up to that time point, to simulate a "real-time" outpatient environment. The provider was asked whether, based on his/her clinical expertise, the patient would be judged as having (a) worsened ("the patient's condition is sufficiently worsened to be considered at higher risk for seizures than last month"), (b) improved ("the patient's condition is sufficiently improved to be considered at lower risk for seizures than last month"), or (c) remained the same ("no change in the patient's condition; the risk for seizures is the same as it was last month") in underlying seizure risk compared to the prior month. Once a response was submitted, the provider was advanced to the following month, and the process was repeated. To avoid confounding interpretation of provider responses, providers were instructed to assume all other factors remained equal from month to month, including treatment and distribution of seizure types. Providers were provided additionally with a "historical database" composed of seizure diaries from 100 similar patients drawn from the SeizureTracker.com database (for real cases), or 100 generated patients (for synthetic cases). A sample of diary entries shown to each provider is shown in Figure 1. The same patient cases were repeated 4 to 12 weeks after the initial questionnaire.

Providers' responses were compared to EpiSAT predictions using the same set of patient cases. For each patient case, EpiSAT was trained on the same set of data as shown to each provider: 100 seizure diaries from the "historical database" and seizure frequencies only up to each revealed diary entry. The number of latent states (K) was estimated based on the value of K minimizing the Bayesian information criterion (BIC) over a grid of latent states, $K = (2,3,4)$, on the historical database. For each outpatient visit (seizure diary entry), 5000 iterations and 2000 sweeps of burn-in (samples prior to convergence) were used. The posterior mode of the latent state at time t (majority value of posterior state estimates across MCMC samples) was used as the EpiSAT estimate of the seizure risk. Seizure risk was classified as worse, better, or unchanged from the previous time point by comparing the posterior mode of the seizure risk level at time t to that of time $t-1$.

2.4 | Real seizure diaries (seizure risk unknown)

This section describes the real seizure diaries, where true seizure risk is unknown, presented to clinicians and EpiSAT in the above prospective testing format. Four real seizure diaries were randomly selected from the SeizureTracker.com database (Table 1). To provide reasonable sample size for calculating provider and EpiSAT accuracy, inclusion criteria required at least 30 seizure diary entries per diary (30 seizure diaries entries per diary x 4 diaries = total 120 seizure entries). The same set of real diaries was tested on clinicians and EpiSAT.

TABLE 1 Characteristics of patient seizure diaries

		Age at initial diary entry (years)	Sex	Monthly seizure frequency (seizures per month)	
				Mean (SD)	Median (MAD)
Synthetic diaries	Patient 1	3	Male	13.5 (22.8)	3.0 (11.5)
	Patient 2	22	Female	10.3 (12.9)	6.0 (8.3)
	Patient 3	8	Male	8.9 (15.2)	3.0 (7.0)
	Patient 4	8	Male	12.6 (31.0)	4.0 (8.6)
Real diaries	Patient 5	1	Male	10.1 (19.4)	4.0 (8.1)
	Patient 6	32	Male	11.8 (9.6)	9.0 (6.8)
	Patient 7	22	Female	6.3 (7.0)	4.0 (3.7)
	Patient 8	8	Male	5.6 (3.9)	5.0 (3.4)

Note: Age at initial diary entry, sex, and monthly seizure frequency of each case are shown.

Cases were randomly sampled from patients using SeizureTracker.com for electronic seizure diary recordings.

Abbreviations: MAD, median absolute deviation; SD, standard deviation.

2.5 | Synthetic seizure diaries (seizure risk known)

This section describes the generation of synthetic seizure diaries, where true seizure risk is known and therefore accuracy of seizure risk estimation can be calculated. Real diaries have unknown true risk, which does not allow for accuracy verification in such data. Both real and synthetic diaries were examined to evaluate similarity to clinician decision patterns across both real and synthetic settings. A two-stage approach was used to generate synthetic seizure diaries and presented in the above prospective testing format to clinicians and EpiSAT (Figure S3):

Stage 1: "Known" latent seizure risk states were first fitted through a nonhomogeneous three-state first-order hidden Markov model with a zero-inflated Poisson emission to 101 randomly sampled diaries from the SeizureTracker.com database. This allows generation of "known" latent states that maintain the temporal autocorrelation structure present in real seizure diaries. Covariates included sex, age, average seizure duration in the prior month, and number of generalized motor seizures. For patient privacy, a discrete uniform (0,20) random variate was added to each age, and patient sex was randomly drawn from a Bernoulli

distribution with proportions according to gender distribution in SeizureTracker.com. After generation of latent risk states, original seizure frequencies themselves were not further used in order to avoid model contamination of the tested data. Time-varying latent seizure risk levels were then used as "ground truth."

Stage 2: Conditional on "known" seizure risk levels, seizure frequencies were generated according to a zero-inflated negative binomial distribution, which has been found to generate realistic clinical seizure diary data.[16] The level of zero-inflation and overdispersion was specified to span the range of empirically observed values for seizure diary data, including zero-inflation of 10^{-6} and dispersion of 0.7-1.1.[16]

This process is necessary to evaluate accuracy in classification problems involving latent variables that cannot be directly measured. This simulation scheme has several notable attributes: (a) temporal autocorrelation structure is maintained, (b) the scheme intentionally assumes that the EpiSAT algorithm is incorrectly specified, and (c) the scheme uses a larger set of covariates to generate seizure risk than that shown to EpiSAT or providers, to simulate a situation in which providers and the EpiSAT algorithm operate under incomplete information about factors influencing seizure risk. Of 101 generated synthetic seizure diaries with seizure risk known, four diaries were randomly selected, each with 30 seizure diary entries (total 120 diary entries). The same set of synthetic diaries was tested on clinicians and EpiSAT.

2.6 | Statistical methods

Performance was evaluated using accuracy, interrater reliability, and intrarater reliability. Accuracy of seizure risk evaluation was calculated as the percentage of correctly identified increases, decreases, or unchanged seizure risk states using the four synthetic seizure diaries (120 seizure entries) for which seizure risk was known. Density-based spatial clustering for applications with noise (DBSCAN) was used to identify outlying epilepsy provider accuracies. Significant differences between clusters in patient volume, years in epilepsy practice, and consideration of seizure risk in clinical practice were evaluated with a Fisher exact test (categorical variables) and Mann-Whitney *U* (continuous variables) with false discovery rate control at the 0.05 level.[17] A two-sample test of proportions was used to evaluate whether a significant difference in accuracy was present between epilepsy providers and EpiSAT. Krippendorff's α was used to evaluate human interrater reliability (ie, between providers) accounting for missingness, calculated on the complete set of eight synthetic and real seizure diaries (240 seizure diary entries). Values of α range from 0 to 1, where 0 indicates perfect disagreement and 1 indicates perfect agreement (Appendix S2). Scott's π was used to evaluate human intrarater reliability (ie, longitudinal reliability within each provider), by comparing the initial set of eight synthetic and real seizure diaries to the repeated evaluation 4 to 12 weeks after the first set of cases. The level of agreement between clinicians and EpiSAT was estimated using the proportion of observed agreement for real and synthetic diaries. Statistical significance was evaluated at the 0.05 level. Consistency of EpiSAT across algorithm initializations was evaluated over 20 random seed initializations. Statistical analyses were performed using R version 3.5.0.

3 | RESULTS

3.1 | Demographics

A total of 24 epilepsy providers were evaluated from the University of California, San Francisco; Beth Israel Deaconess Medical Center; and Baylor College of Medicine, including 12 ABPN epilepsy board-certified epileptologists (50.0%), 4 epilepsy nurse practitioners (16.7%), and 8 epilepsy fellows (33.3%). The majority of providers (56.5%) saw more than 20 epilepsy patients per month and had been practicing for a mean of 9.3 years (standard deviation [SD] 10.2) postresidency in epilepsy clinical practice (Figure 2). Seventeen of 24 providers (70.8%) reevaluated seizure diaries a second time as part of intrarater reliability assessment. Twenty-one of 24 providers (87.5%) stated that they actively attempt to distinguish seizure risk from natural fluctuations in seizure count in clinical practice. The mean interval between assessments was 7.2 weeks (SD 2.3 weeks) and ranged from 4.0 to 11.9 weeks. Characteristics of patient diaries are shown in Table 1.

3.2 | Epilepsy provider accuracy in seizure risk assessment

The mean accuracy of clinicians for correctly identifying whether seizure risk had improved, worsened, or remained unchanged based on seizure counting data was 74.7% (SD 13.6%). Spatial clustering identified a bimodal distribution of accuracies in seizure risk evaluation: a dominant mode of providers (n = 20, Figure 3A, circles) and a secondary mode of providers (n = 4, Figure 3A, triangles). Among the dominant cluster, a monotonic increasing relationship between clinical epilepsy experience and accuracy in identifying seizure risk changes was present (Spearman's $\rho = 0.56$; $P = .01$), with the relationship leveling off after approximately 20 years (Figure 3B). Among the secondary mode of four providers, there was no association between

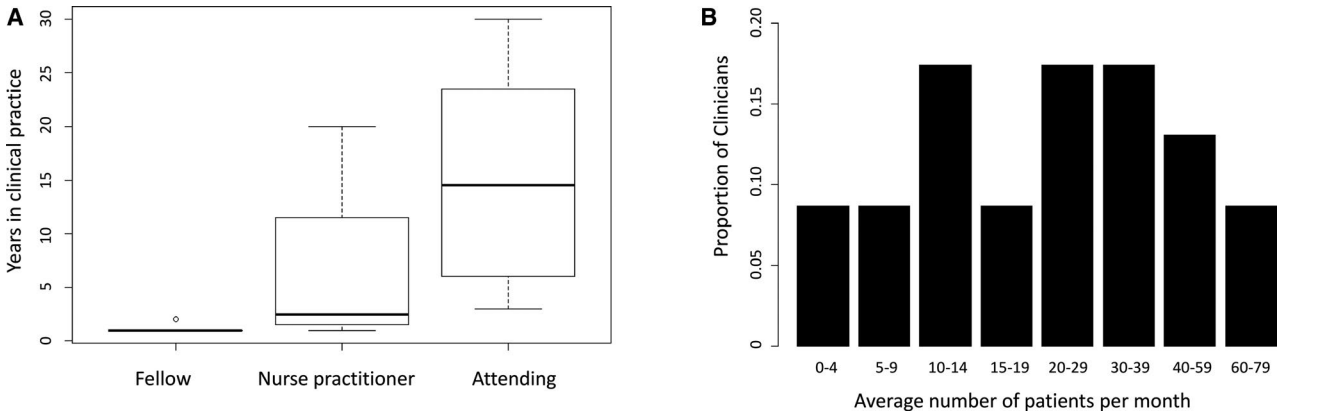

FIGURE 2 Clinical experience of study participants. A, Epilepsy providers had on average 9.3 years of clinical experience (SD 10.2). B, All epilepsy providers were in active clinical practice, with the majority seeing more than 20 patients per month

clinical experience and accuracy (Figure 3C), although with limited interpretability due to the small sample size (n = 4). These four providers did not exhibit a clear difference from the dominant cluster with regard to the proportion considering seizure risk in clinical practice ($P > .99$), current patient volume ($P > .99$), or years of clinical epilepsy experience ($P = .08$). The four providers exhibited a decision-making pattern that was different from the other 20 providers in that the secondary mode tended to underrecognize worsening or improvement in seizure risk (Table S1, Clinicians 21-24).

3.3 | EpiSAT accuracy in seizure risk assessment

A substantial proportion of observed agreement was present between clinicians and EpiSAT in synthetic seizure diaries (mean 75.4%, SD 12.6%). As shown in the example in Figure 4, the majority decision by clinicians was commonly compatible with EpiSAT predictions (Figure 4B-C). Real seizure diaries yielded a similar proportion of observed agreement between clinicians and EpiSAT (mean 71.9%, SD 13.5%). Mean accuracy of EpiSAT for seizure risk estimation was 87.5% (SD, 5.7%), yielding a 17.4% significantly higher proportion of correctly identified seizure risk changes using EpiSAT compared to providers (two-sample test of proportions, $P = .002$) (Table 2). EpiSAT obtained higher accuracy for identifying seizure risk changes than 21 of 24 providers (Table S1). Posterior mode estimates were robust across initializations.

3.4 | Interrater and intrarater reliability of seizure risk assessment

Low-to-moderate interrater reliability (Krippendorff's $\alpha = 0.45$) was present between clinicians for seizure risk identification. Although the decision with the highest majority across clinicians was often congruent with the correct assessment of seizure risk change, there was substantial variability between providers in identifying whether a change in seizure frequency reflected worsening, improvement, or no change in seizure risk (Figure 4B). Intrarater reliability was reasonably consistent within each clinician (Scott's $\pi = 0.89$) but ranged between 0.61 and 1.0 depending on the provider. There was no significant association between increasing duration of time between assessments and intrarater reliability (Pearson $r = 0.09$, $P = .74$) (Figure S4).

4 | DISCUSSION

We conducted initial clinical testing of a Bayesian machine learning algorithm, EpiSAT, as a potential clinical decision support tool for quantitative analysis of changes in clinical seizure frequency, and we compared performance in a prospective study design to specialized epilepsy providers at three academic NAEC level 4 epilepsy centers in the United States. Among surveyed clinicians, the majority reported distinguishing clinically between a concept of seizure risk and seizure frequency in clinical decisions. However, we show that even epilepsy experts exhibit only low-to-moderate consistency in using raw clinical seizure frequencies to judge underlying seizure risk, with a large degree of variability between providers. We found that EpiSAT exhibits substantial observed agreement with human experts, supporting the clinical validity of seizure risk recommendations made by EpiSAT. We furthermore show that EpiSAT may yield improved accuracy for seizure risk detection and is highly consistent in a clinically realistic controlled setting. Improved consistency of care may help improve patient outcomes and reduce adverse events.

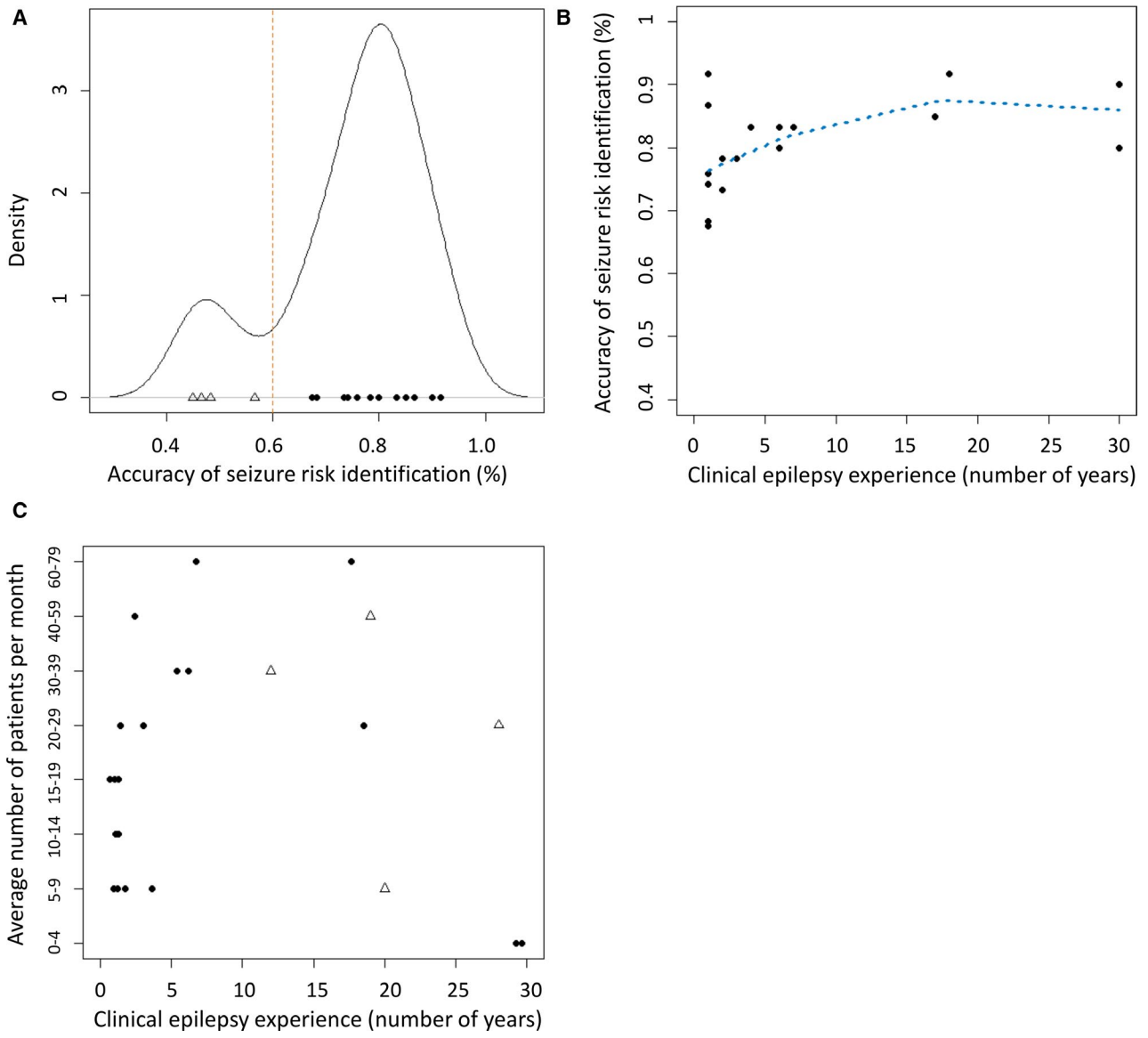

FIGURE 3 Accuracy of epilepsy providers in identifying seizure risk. A, Two distinct groups of providers were identified based on DBSCAN clustering of accuracy for assessing risk changes: a main group (n = 20, circles) and a secondary outlying group (n = 4, triangles). B, Clinical experience was associated with accuracy of seizure risk identification in the main group of providers, which leveled off after about 20 years (Spearman $\rho = 0.56$; $P = .01$). Cubic smoothing spline with 3 degrees of freedom is superimposed in blue. C, Differences in experience and patient volume did not clearly separate the two groups of providers

There are several main contributions of our work. First, our results show that the use of EpiSAT is potentially useful as a clinical decision support tool for quantitative analysis of breakthrough seizures and changes in seizure frequency and yields recommendations similar to those of epilepsy experts but with higher accuracy and reliability. EpiSAT is designed to mathematically distinguish changes in risk versus natural variation in seizure frequencies. Using a tool to systematically identify patients' degree of "seizure control" rather than subjective evaluation of raw seizure frequencies may facilitate more appropriate responses to apparent fluctuations in reported seizures, which may be confounded by missed seizure counting and natural variability. This concept of distinguishing seizure control as a separate entity from raw seizure frequencies was found consistent with how the majority of epilepsy experts practice. However, no systematic approach currently exists for this task. The EpiSAT system provides a possible clinical decision support tool to address this need. EpiSAT exhibited substantial observed agreement with clinicians, demonstrating that EpiSAT exhibits decision-making patterns consistent with epilepsy experts and supporting the clinical validity of seizure risk recommendations made by EpiSAT. EpiSAT attained more accurate seizure risk identification than 21 of 24 clinicians, with average EpiSAT accuracy yielding a 17.6% improvement over average clinician accuracy. This accuracy is likely to further improve

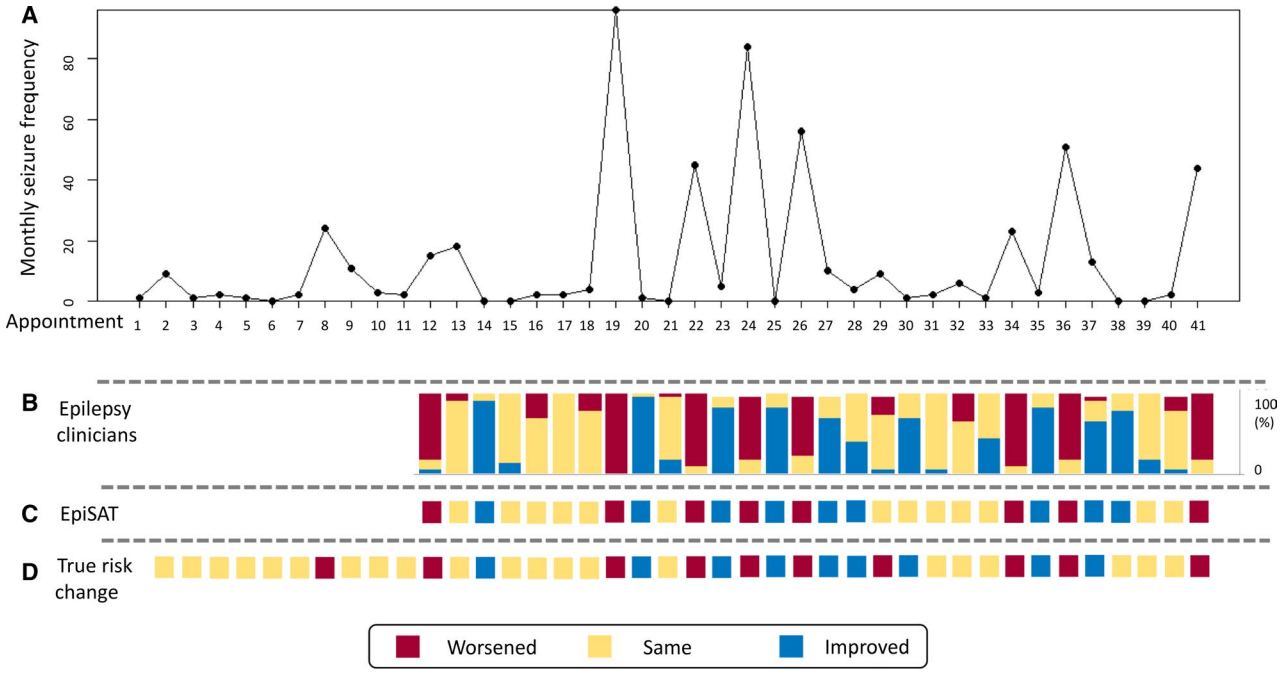

FIGURE 4 Seizure diary entries and comparison of clinician versus EpiSAT decisions for Case 1. A, Patient-reported seizure frequencies at each appointment. B, Distribution of clinician judgments of seizure risk improvement (blue), worsening (red), or lack of change (yellow) from prior visit. The length of the bar indicates the percentage of clinicians with each response. C, EpiSAT estimate of seizure risk change, based on posterior mode of latent state. D, True seizure risk change from underlying generative process

TABLE 2 Clinician and EpiSAT performance for seizure risk identification on the four synthetic diaries

	Number of seizure diary entries	Clinicians (%)	EpiSAT (%)
Patient 1	30	Mean (SD): 76.1% (13.4%) Median (MAD): 80.0% (4.0%) (n = 23)	90.0
Patient 2	30	Mean: 72.5% (18.4%) Median: 80.0% (7.5%) (n = 24)	86.7
Patient 3	30	Mean: 73.3% (13.2%) Median: 75.0% (1.7%) (n = 6)	93.3
Patient 4	30	Mean: 76.1% (9.3%) Median: 80.0% (3.9%) (n = 6)	80.0

Note: Proportion of correct risk identifications is reported.

Each patient case consisted of 30 seizure diary entries.

EpiSAT obtained higher accuracy for all individual patient cases as well as overall.

Abbreviations: MAD, median absolute deviation; SD, standard deviation.

in patients with greater periodicity of seizure rhythms. These results suggest potential clinical utility of EpiSAT as augmented intelligence within a clinical decision support system.

Second, the results of our study suggest that human performance for evaluating seizure risk based on clinical seizure diaries is relatively inconsistent across providers and has low-to-moderate reproducibility in an artificial setting. Currently, there is no systematic methodologic approach employed in neurology settings for objective determination of when breakthrough seizures or apparent changes in clinical seizure frequency are likely to reflect a true change in underlying seizure propensity, versus natural variation. As demonstrated here, even expert opinion can vary highly in assessment of when increases in raw seizure frequency should be judged as worsened seizure propensity. Human assessment may vary depending on many factors, including differences in training, fatigue, time constraints, diagnostic biases, or distractibility. Systematic algorithms have become of interest for augmented intelligence, as they rely on objective algorithms to minimize subjectivity in evaluation, and furthermore are not prone to human factors. The interrater reliability of clinicians in our sample was low-to-moderate, and similar to the interrater reliability present between epileptologists for reading EEG and chronic ambulatory electrocorticography.[18,19] Low interrater reliability between clinicians may result in variability in patient outcomes from differing management recommendations. Although intervention decisions are often complex assessments based on multiple factors including seizure frequency, adverse effects, and quality of life, improved reproducibility using EpiSAT has the potential to complement the clinical workflow and improve consistency of care.

As expected, each clinician was largely consistent within him/herself when asked to reevaluate the same patient after 4 to 12 weeks, although we found that clinicians exhibited a range of internal consistencies ($\pi = 0.61$-1.0).

One of the ethical challenges that has emerged with the rise of artificial or augmented intelligence is the "black box issue," given ethical concerns of employing systems in which it is not known how the system derives its actions when decisions may inform or guide clinical decisions. A major benefit of the EpiSAT software is the Bayesian model–based approach to seizure risk estimation. Because Bayesian algorithms are built using hierarchical modeling in which a transparent set of variables can be evaluated, they are considered to be more transparent than approaches that do not contain the opacity that may be present with other machine-learning techniques. Specifically, the "inner workings" of the model are clearly specified based on the modeled relationships between the latent variable capturing states of heightened seizure susceptibility, observed clinical seizures, temporal dependency, and baseline patient characteristics. The use of transparent, model-based approaches to seizure risk identification can help to ensure safety in augmented intelligence applications. Other state-space models may also be of interest to explore; an overview of alternative methods is provided in Appendix S3.

This study demonstrates feasibility of use of EpiSAT as a clinical decision support system to guide clinical interpretation of breakthrough seizures and changes in seizure frequency. Applications of EpiSAT may be targeted toward the clinical question; for example, for patients with multiple seizure types, EpiSAT may be run separately on each seizure type to evaluate the evolution of each type. Indeed, many patients with seizures are treated in general neurology practice and not always referred to an epilepsy specialist. As a clinical decision support system, EpiSAT can offer guidance that parallels decision-making by clinical experts in general neurology or resource-limited settings. We found that, for the majority of epilepsy experts, more clinical experience was associated with higher accuracy in assessing when seizure count fluctuations reflected real changes in seizure burden, suggesting that more clinical experience in epilepsy may lead to better ability to differentiate changes in seizure burden from natural fluctuation. For a minority, more years in practice did not clearly lead to higher accuracy of seizure risk identification; these clinicians were those who tended to underidentify seizure risk changes. A recent study on 30-day mortality among high-risk patients with heart failure or cardiac arrest was lower during the dates of national cardiology meetings; although the rationale for this observed trend is unclear, one possibility is that the composition of physicians who remain in house during national conferences is different, which in many institutions may include more fellows than faculty.[18] Further data are needed with denser sampling across various levels of experience to further evaluate this hypothesis. Due to the ability to process massive amounts of information, clinical decision support systems have the capability of identifying patterns that would otherwise require years of experience by human learning or may even detect subtle patterns that human experts may not be able to detect. EpiSAT also consumes less clinician time than diary review, which allows more time for patient counseling or education. With the development of new noninvasive seizure detection devices and the increasing accessibility of electronic seizure diary user interfaces, we expect that the reliability of data collected by electronic clinical seizure diaries will only continue to improve.

There are several limitations to this validation study. (a) Because mandatory participation was not required, nonresponse bias may result in comparison to a group of epileptologists who are a priori interested in understanding seizure risk. Future work enrolling larger, prospective clinician arms with different degrees of expertise, including general neurologists and primary care physicians, is the next expected stage. (b) Inability to include all NAEC level 4 epilepsy centers in the United States may lead to selection bias—sampling primarily academic epileptologists from a small percentage of NAEC level 4 epilepsy centers—and may limit study generalizability. (c) To enhance responsiveness, the number of cases per provider was purposefully kept short, and future work ought to evaluate performance across multiple epilepsy etiologies, periodicities, and seizure frequencies. (d) Finally,, it may be argued that given that EpiSAT is built on a model of seizure risk, it is not surprising that EpiSAT outperforms the average epileptologist in identifying changes in underlying seizure risk. However, the majority of clinicians reported actively distinguishing seizure risk from natural fluctuation in clinical practice. Furthermore, the majority decision from clinicians was often consistent with the underlying generative seizure risk changes, supporting the validity of EpiSAT's recommendations.

The results validate the ability of EpiSAT to yield decision-making recommendations overall similar to specialized epilepsy providers, but with higher accuracy and reproducibility than current practice. Use of EpiSAT as a clinical decision support tool allows for a systematic approach to seizure risk, which may help improve patient management, as well as streamline and improve epilepsy care.

ACKNOWLEDGMENTS

This study was approved by the Office of Human Subject Research Protection under protocol #12301. Use of the SeizureTracker.com data was facilitated by the International Seizure Diary Consortium (https://sites.google.com/site/isdchome/). Deidentified seizure diaries were exported

from SeizureTracker.com in accordance with NIH OHSRP Protocol #12301. We thank the SeizureTracker.com patients for donating their seizure diaries and epilepsy providers at the University of California, San Francisco, Beth Israel Deaconess Medical Center, and Baylor College of Medicine for donating their time and participation.

CONFLICT OF INTERESTS

R.M. is the cofounder/owner of Seizure Tracker, LLC, and reports personal fees from Cyberonics, Courtagen, Engage Therapeutics, Neurelis, UCB, Brain Sentinel; and grants from the Tuberous Sclerosis Alliance. D.G. is on the medical advisory board of Magic Leap and has received grant support from BIDMC and NIH T32NS048005. W.H.T. reports grant support from NIH ZIANS002236. M.V. reports partial support from NSF SES-1659925 and NSF DMS-1811568. The remaining authors have no conflicts of interest to report. 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

Sharon Chiang https://orcid.org/0000-0002-4548-4550
Daniel M. Goldenholz https://orcid.org/0000-0002-8370-2758
Vikram R. Rao https://orcid.org/0000-0002-6389-2638
William H. Theodore https://orcid.org/0000-0002-4669-5747
Jonathan K. Kleen https://orcid.org/0000-0003-2622-3205
Jay Gavvala https://orcid.org/0000-0002-9392-6608
Marina Vannucci https://orcid.org/0000-0002-7360-5321
John M. Stern https://orcid.org/0000-0002-3549-1642

REFERENCES

1. Bone I, Fuller GN. Neurology in practice: epilepsy. J Neurol Neurosurg Psychiatry. 2001;70(Suppl 2):Ii1-2.
2. Stone J, Carson A, Duncan R, Roberts R, Warlow C, Hibberd C, et al. Who is referred to neurology clinics?–the diagnoses made in 3781 new patients. Clin Neurol Neurosurg. 2010;112:747–51.
3. Wile DJ, Warner J, Murphy W, Lafontaine AL, Hanson A, Furtado S. Referrals, wait times and diagnoses at an urgent neurology clinic over 10 years. Can J Neurol Sci. 2014;41:260–4.
4. Karoly P, Goldenholz DM, Cook M. Are the days of counting seizures numbered? Curr Opin Neurol. 2018.
5. Cook B, Cook M. Does my posterior look big in this? Bayesian solutions to seizure counting problems. Epilepsia Open. 2019;4:235–6.
6. Goldenholz DM, Goldenholz SR, Moss R, French J, Lowenstein D, Kuzniecky R, et al. Is seizure frequency variance a predictable quantity? Ann Clin Transl Neurol. 2018;5:201–7.
7. Goldenholz DM, Moss R, Scott J, Auh S, Theodore WH. Confusing placebo effect with natural history in epilepsy: a big data approach. Ann Neurol. 2015;78:329–36.
8. Goldenholz DM, Strashny A, Cook M, Moss R, Theodore WH. A multi-dataset time-reversal approach to clinical trial placebo response and the relationship to natural variability in epilepsy. Seizure. 2017;53:31–6.
9. Baud M, Schindler K. Forecasting seizures: not unthinkable anymore.
10. Baud MO, Kleen JK, Mirro EA, Andrechak JC, King-Stephens D, Chang EF, et al. Multi-day rhythms modulate seizure risk in epilepsy. Nat Commun. 2018;9:88.
11. Freestone DR, Karoly PJ, Peterson AD, Kuhlmann L, Lai A, Goodarzy F, et al. Seizure prediction: science fiction or soon to become reality? Curr Neurol Neurosci Rep. 2015;15:73.
12. Karoly PJ, Goldenholz DM, Freestone DR, Moss RE, Grayden DB, Theodore WH, et al. Circadian and circaseptan rhythms in human epilepsy: a retrospective cohort study. Lancet Neurol. 2018;17(11):977–985.
13. Spencer DC, Quigg MS, Fountain NB, Jobst BC, Wong VS, Mirro E, et al. Interrater reliability in interpretation of electrocorticographic seizure detection of the responsive neurostimulator. Epilepsia. 2018;56(6):968–971.
14. Chiang S, Vannucci M, Goldenholz DM, Moss R, Stern JM. Epilepsy as a dynamic disease: a Bayesian model for differentiating seizure risk from natural variability. Epilepsia Open. 2018;3:236–46.
15. Shorvon SD, Perucca E, Engel J. The Treatment of Epilepsy. 4th edn. Oxford: Wiley-Blackwell, 2015.
16. Tharayil JJ, Chiang S, Moss R, Stern JM, Theodore WH, Goldenholz DM. A big data approach to the development of mixed-effects models for seizure count data. Epilepsia. 2017;58:835–44.
17. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B. 1995;57:289–300.
18. 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.
19. Quigg M, Sun F, Fountain NB, Jobst BC, Wong VS, Mirro E, et al. Interrater reliability in interpretation of electrocorticographic seizure detections of the responsive neurostimulator. Epilepsia. 2015;56:968–71.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Chiang S, Goldenholz DM, Moss R, et al. Prospective validation study of an epilepsy seizure risk system for outpatient evaluation. *Epilepsia*. 2020;61:29–38. https://doi.org/10.1111/epi.16397