




RESEARCH ARTICLE

Estimating the likelihood of epilepsy from clinically noncontributory electroencephalograms using computational analysis: A retrospective, multisite case–control study

Luke Tait^{1,2} | Lydia E. Staniaszek^{3,4} | Elizabeth Galizia⁵ | David Martin-Lopez^{5,6} | Matthew C. Walker^{7,8}  | Al Anzari Abdul Azeez⁸ | Kay Meiklejohn^{4,9} | David Allen⁹ | Chris Price¹⁰ | Sophie Georgiou¹⁰ | Manny Bagary¹¹ | Sakh Khalsa¹¹ | Francesco Manfredonia¹² | Phil Tittensor^{12,13} | Charlotte Lawthom^{14,15} | Benjamin B. Howes⁴ | Rohit Shankar^{16,17}  | John R. Terry^{2,4} | Wessel Woldman^{2,4} 

Correspondence

Wessel Woldman, University of Birmingham, Birmingham, UK.
Email: w.woldman@bham.ac.uk

Funding information

Innovate UK, Grant/Award Number: 103939; Engineering and Physical Sciences Research Council, Grant/Award Number: EP/N014391/2 and EP/T027703/1; National Institute for Health and Care Research, Grant/Award Number: AI01646; Epilepsy Research UK, Grant/Award Number: F2002

Abstract

Objective: This study was undertaken to validate a set of candidate biomarkers of seizure susceptibility in a retrospective, multisite case–control study, and to determine the robustness of these biomarkers derived from routinely collected electroencephalography (EEG) within a large cohort (both epilepsy and common alternative conditions such as nonepileptic attack disorder).

Methods: The database consisted of 814 EEG recordings from 648 subjects, collected from eight National Health Service sites across the UK. Clinically noncontributory EEG recordings were identified by an experienced clinical scientist ($N = 281$; 152 alternative conditions, 129 epilepsy). Eight computational markers (spectral [$n = 2$], network-based [$n = 4$], and model-based [$n = 2$]) were calculated within each recording. Ensemble-based classifiers were developed using a two-tier cross-validation approach. We used standard regression methods to assess whether potential confounding variables (e.g., age, gender, treatment status, comorbidity) impacted model performance.

Results: We found levels of balanced accuracy of 68% across the cohort with clinically noncontributory normal EEGs (sensitivity = 61%, specificity = 75%, positive predictive value = 55%, negative predictive value = 79%, diagnostic odds ratio = 4.64, area under receiver operated characteristics curve = .72). Group level

Luke Tait and Lydia E. Staniaszek are joint first authors.

Rohit Shankar, John R. Terry, and Wessel Woldman are joint senior authors.

For affiliations refer to page 10.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Epilepsia* published by Wiley Periodicals LLC on behalf of International League Against Epilepsy.

analysis found no evidence suggesting any of the potential confounding variables significantly impacted the overall performance.

Significance: These results provide evidence that the set of biomarkers could provide additional value to clinical decision-making, providing the foundation for a decision support tool that could reduce diagnostic delay and misdiagnosis rates. Future work should therefore assess the change in diagnostic yield and time to diagnosis when utilizing these biomarkers in carefully designed prospective studies.

KEYWORDS

biomarker, case-control, computational, EEG, network

1 | INTRODUCTION

Epilepsy affects >50 million people worldwide, with estimates suggesting approximately 2.4 million new cases worldwide per year.¹ Epilepsy remains a clinical diagnosis, based on expert analysis of likelihood of further seizures, a decision that considers multiple factors including a person's medical history and results from routine diagnostic tests such as the scalp electroencephalography (EEG). However, the sensitivity of routine EEG in the identification of persons with epilepsy remains low, relying on expert identification of interictal epileptiform discharges (IEDs).^{2,3} The overall specificity of routine EEG is broadly estimated to fall within a range of 78%–98%, but individual studies often report wide confidence intervals (CIs).^{2,3} Additionally, it is currently recommended that EEGs lacking IEDs (herein termed “noncontributory EEGs”) should not be used in isolation to exclude a diagnosis of epilepsy (see e.g. NICE guidelines 2017⁴). As a result, delay in both the diagnosis of epilepsy and its differentials is common, driving research into identification of biomarkers of epilepsy, using routine EEG as a substrate. Computational approaches to interrogate routine EEG have attracted much interest recently. Typically, these have focused on automatic identification of IED, and/or computation of whole-brain networks by analysis of resting-state EEG, that is, those portions of EEG in which no epileptiform features are present.^{5–7} Automatic identification of IED is both sensitive to EEG artifacts and limited in its performance by the levels of identification achieved by the trained expert. Recently, AI approaches such as SCORE-AI have been proposed for the analysis of routine EEGs, with a particular focus on the identification of (diagnostic) abnormalities.^{8,9}

In contrast, approaches that do not require the presence of IED offer the potential to improve sensitivity of routine EEG in a complementary fashion to present clinical practice. Here, the focus is on features that are not

Key points

- Analysis of resting-state EEG has revealed group level differences between people with epilepsy and those with an alternative condition.
- We validated an existing set of eight biomarkers (spectral-, network-, and model-based) in a representative population of clinically noninformative EEGs ($N=281$).
- Statistical classifiers trained on these noninformative EEGs showed better-than-chance performance: sensitivity of 61% and diagnostic odds ratio of 4.64 by RUSboost.
- The study findings demonstrate the potential added value of computational biomarkers from EEG for people with suspected epilepsy and seizures.
- By offering decision support in clinically noninformative EEG, these methods might in the future contribute to reduced diagnostic delay and misdiagnosis rates.

currently considered clinically informative. For example, computational analysis of resting-state EEG has consistently revealed differences in whole-brain network measures at the group level. These differences, confirmed by meta-analysis, have been shown in case-control studies in both generalized and focal epilepsies when compared to healthy participants.^{6,10} Although constituting phase 1 level evidence, it is unclear whether these findings are translatable to patient cohorts typical of those seen in clinical practice. Few studies have assessed group level changes in resting-state or clinically noncontributory EEG between persons with epilepsy and persons with an alternative condition (e.g., syncope or nonepileptic attack disorder [NEAD]) who may also be referred for an EEG

following a seizurelike event. However, those that did have also confirmed the presence of identifiable group level differences, demonstrating the potential of computational biomarkers both in the presence of identifiable IED and in noncontributory EEGs.¹¹ It is also unclear how group level effects translate to the individual level. For example, the group level differences observed in Larsson and Kostov¹² were shown to offer very limited predictive capacity at the individual level.^{5,13} A further limitation in the studies described above is that typically they were performed on data obtained from a small number of diagnostic centers.

Addressing these challenges, we developed a robust classification pipeline to validate a set of candidate biomarkers in a way that is relevant to clinical practice. To this end, we performed a retrospective, multisite case-control study. Our study includes people ultimately diagnosed with epilepsy, as well as those ultimately diagnosed with common alternatives including syncope and NEAD. To maximize robustness and increase its clinical applicability, we leveraged a two-tier cross-validation approach, and made no exclusions based on comorbidity (including neurological), medication (including antiseizure medications or medications with known effects on the EEG), or time since first, or most recent, seizure or seizurelike event.

2 | MATERIALS AND METHODS

2.1 | Study design and participants

Eight sites within the National Health Service (NHS) participated in this study (Figure 1). Inclusion and

exclusion criteria may be found at clinicaltrials.gov (identifier: NCT05384782). In summary, inclusion requirements were as follows: adult; suspected of having had a seizure or epilepsy; one or more EEGs recorded as part of the diagnostic process with a minimum of 19 channels, applied to the 10–20 international system of electrode placement; and a final diagnosis of epilepsy or a common alternative condition (e.g., syncope, non-epileptic attack disorder) determined by a clinical expert (e.g., neurologist), which has remained stable for at least 1 year. To minimize selection bias, participants who met the inclusion criteria were included in backward chronological order. An initial target total of 100 participants was set for each participating site. NHS sites supplied EEG traces alongside metadata detailing patient sex, age, comorbidities, medication status, EEG result (normal, abnormal but not diagnostic, or diagnostic of epilepsy), and final diagnosis (Table 1).

EEGs that were reported by the consultant neurophysiologist of each participating site as “normal” or “abnormal” (but without diagnostic features) were included in this study. For classification purposes, an abnormal EEG included EEGs that contained abnormal features that were not specific for epilepsy, and that therefore did not contribute to the ultimately confirmed diagnosis. An example of such an EEG could include incidental findings of nonspecific abnormalities such as those that may be of a vascular, pharmacological, structural, or metabolic pathophysiological origin. Due to the inherent heterogeneity introduced by the presence or absence of EEG abnormalities, individual classifiers will be developed for the normal and the abnormal clinically noncontributory

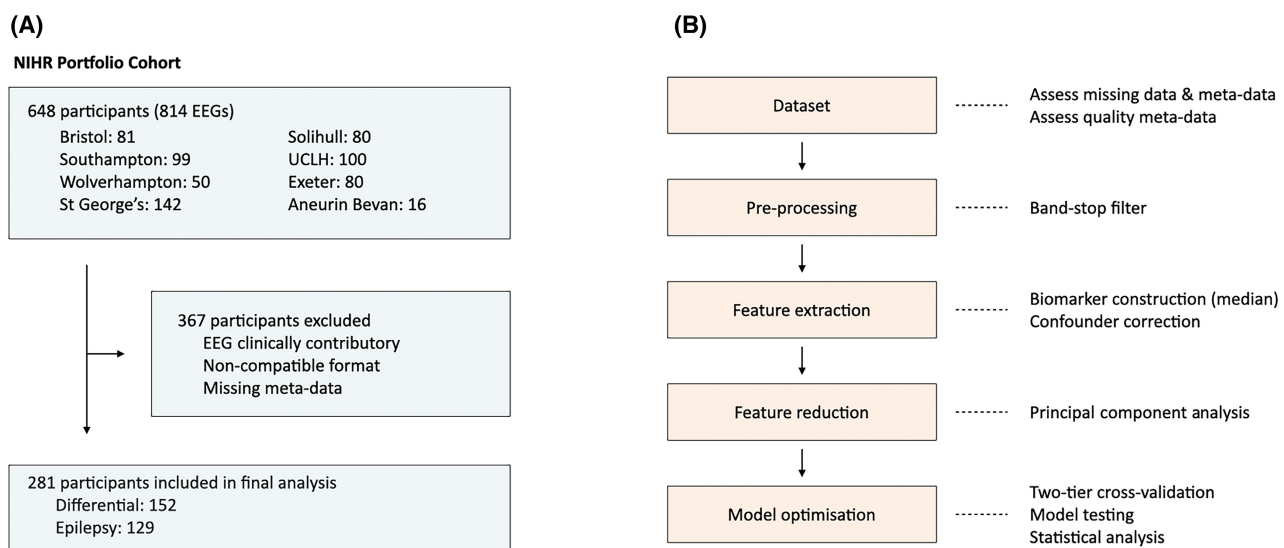


FIGURE 1 Cohort profile and analysis pipeline. (A) Cohort profile. Final analysis was carried out on the first available electroencephalogram (EEG) of each participant that was clinically noncontributory. (B) Flow diagram summarizing analysis steps. UCLH, University College London Hospitals.

Characteristic	Abnormal EEG	Normal EEG	Total
Age, mean (SD; range), years	50.5 (20.1; 18–91)	38.2 (14.4; 18–91)	42.0 (17.4; 18–91)
Sex, <i>n</i> (%)			
Female	58 (65.2)	102 (53.1)	160 (56.9)
Male	31 (34.8)	89 (46.4)	120 (42.7)
Other	0 (0)	1 (.5)	1 (.4)
ASM treatment status, <i>n</i> (%)			
Yes	40 (44.9)	70 (36.5)	110 (39.1)
No	49 (55.1)	122 (63.5)	171 (60.9)
Presence of comorbidity, <i>n</i> (%)			
Yes	64 (71.9)	123 (64.1)	187 (66.5)
No	25 (28.1)	69 (35.9)	94 (33.5)

Abbreviations: ASM, antiseizure medication; EEG, electroencephalogram.

cohorts separately. Where more than one EEG was submitted per patient, only the first relevant noncontributory EEG was used for each cohort (this means that subjects with a clinically normal EEG as well as a nonspecific abnormal EEG would have those respective EEGs included in the relevant model arms). In the (anticipated to be rare) case of a dual diagnosis of epilepsy and NEAD (or another common alternative or comorbidity), the subject was included in the epilepsy cohort.

2.2 | EEG preprocessing

Scalp EEG data were imported into MATLAB (R2021b). The same 19 clinical EEG channels (Appendix S1) were chosen for all participants across all sites, whereas any other channels were discarded. Recordings were notch filtered at 50 Hz and bandpass filtered from .53 to 70 Hz using a 4th order Butterworth zero-phase filter and rereferenced to average. Bad channels were interpolated using the Fieldtrip Toolbox.¹⁴ Epochs were rejected if the power within a channel was smaller than 10^{-5} of the median power across channels, or if it contained *z*-scored values > 10. If 33% of the epochs were impacted by such artifacts, a different 20-min window was selected if present, and otherwise the EEG was excluded.

2.3 | Candidate markers

Published resting-state biomarkers of EEG were reviewed in the literature making use of Google Scholar and PubMed. We searched for papers that had reported areas under receiver operating characteristic curve (AUROCs) for statistical models classifying epilepsy from resting-state EEG/magnetoencephalography and used these AUROC values as

TABLE 1 Metadata of the participants included in the study.

the a priori effect size. A total of six papers were identified, with AUROC values ranging from .43 to .99.^{15–20} Following the method of Riley et al.,²¹ we found seven estimates for the maximal number of markers. Of these seven estimates, six suggested a maximum of eight markers and one suggested a maximum of seven markers. Hence, we proceeded to select the first eight markers from an ordered list of candidate markers (ordered by individual effect size).⁵

2.3.1 | Frequency-based markers

Two frequency-based markers were calculated. The peak-alpha frequency was calculated by averaging the power spectrum of the occipital electrodes (O1 and O2) and calculating the frequency corresponding to the peak of the spectrum in the 8–13-Hz range.^{20,22,23} High alpha power was determined by calculating the 10.5–13.5-Hz relative power (averaged across channels after bipolar-referencing of the frontal and temporal channels).¹⁸

2.3.2 | Network-based markers

The low-alpha band (6–9 Hz) phase-locking value (PLV) was computed between pairs of electrodes, and four graph theoretical markers were calculated from these weighted undirected networks.^{24,25} These markers were mean degree, degree variance, average weighted clustering coefficient, and characteristic path length.²⁶

2.3.3 | Model-based markers

The final two markers are derived from the phenomenological multiscaled oscillator model of the EEG.^{13,15}

For both markers, an individual's low-alpha PLV network and alpha power are used to parameterize a model that simulates EEG in which each electrode is described as a system of coupled oscillators (locally coupled according to each channel's alpha power and scaled with a value called "local coupling"), and electrodes are connected according to the PLV weights (scaled by some value called "global coupling"). The first marker, critical coupling, is the theoretical value of global coupling that causes the oscillators to synchronize (i.e., a simulated seizure). The second marker, local coupling, is the theoretical value of local coupling (for a single simulated EEG electrode) that causes synchronization. The marker corresponds to the maximum value across the 19 channels (representing the single most "ictogenic" node in the model).

2.3.4 | Marker calculation

All markers were implemented in MATLAB (R2021b) using built-in MATLAB functions and toolboxes (e.g., the signal-processing toolbox), the brain connectivity toolbox, and previously published scripts for model-based markers.^{13,26} To minimize the effect of artifacts or nonspecific abnormalities, EEG data were segmented into 20-s epochs and the eight markers were calculated for each epoch. The median value for each marker across all 20-s epochs was subsequently used for further analysis. Within each recording, 20 min of EEG was selected. The chosen starting time was sampled from the distribution of starting times of all routine EEGs within the study (Appendix S1).

2.4 | Confounder model

Markers were adjusted for effects of confounders such as age, comorbidities, antiseizure medications (ASM), and sex. To do so, we constructed a generalized linear model of the form:

$$y = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Sex} + \beta_3 \cdot \text{Comorbidity} + \beta_4 \cdot \text{ASM} + \beta_5 \cdot \text{Group} + \varepsilon, \quad (1)$$

where y is a given marker; Age is participant age; Sex, Comorbidity, ASM, and Group each have a value equal to 1 if the participant is respectively male, has comorbid disorders (e.g., dementia, stroke), is being treated with ASM, and has epilepsy or have value zero otherwise; $\beta_i, i \in \{0, \dots, 5\}$ are the parameters of the model; and ε is the participant-specific residuals. Prior to linear modeling, the relevant variables were transformed using a monotonic function (e.g., log

transform) to better approximate normality in the residuals. Note that during hold-out cross-validation, confounder correction was applied to both training and hold-out sets using linear models fitted to the training set (i.e., the β_i estimated from the training set were used for adjusting the markers in the test/hold-out set):

$$y_{\text{adj}} = y - \beta_1 \cdot \text{Age} - \beta_2 \cdot \text{Sex} - \beta_3 \cdot \text{Comorbidity} - \beta_4 \cdot \text{ASM}. \quad (2)$$

2.5 | Statistical model

Statistical models were developed using a two-tiered (or nested) cross-validation scheme (Figure 2). Data were partitioned using nested 10-fold cross-validation, and a suite of statistical models were trained using MATLAB's Classification Learner application (Figure 2). Each individual is part of an *independent* hold-out fold (of size 10%; used to assess the performance of the final model) exactly once; in the remaining nine folds, the individual is part of development sets (of size 90%; approximately 81% of the original size of the entire dataset). A development set is then split into 10 further folds, with one *internal* test set (10% of the development data; approximately 1% of the original dataset) for cross-validation (the remaining 90% of the development data are used for training). In total, 30 different statistical models were assessed during cross-validation (regression-based, decision trees, discriminant analysis, support vector machines, ensemble classifiers, k -nearest neighbors methods, and neural networks; see Appendix S1). For each training fold, principal component analysis (PCA) was performed on the training set (with the aim of feature reduction), and the PCA weights (i.e., hyperparameter) from the training set were subsequently applied to the internal hold-out set. A full description of the hyperparameters for model training is given in Appendix S1. Ultimately, the model with the highest mean cross-validated balanced accuracy was then selected, and its performance was assessed with the independent hold-out folds. In other words, the model is trained 10 times on the development sets (with 10-fold internal cross-validation) and then tested on the 10 corresponding independent hold-out sets. Model performance was characterized using a set of well-established outcome metrics including balanced accuracy, AUROC, and Brier score (Appendix S1). Potential relationships between confounders and overall performance were tested with Fisher exact test (95% significance) and the Kruskal-Wallis test (95% significance), using Bonferroni correction for multiple comparisons. Effect sizes for individual candidate biomarkers were explored using violin plots. The study followed the TRIPOD reporting guideline (reporting guidelines from the equator network).

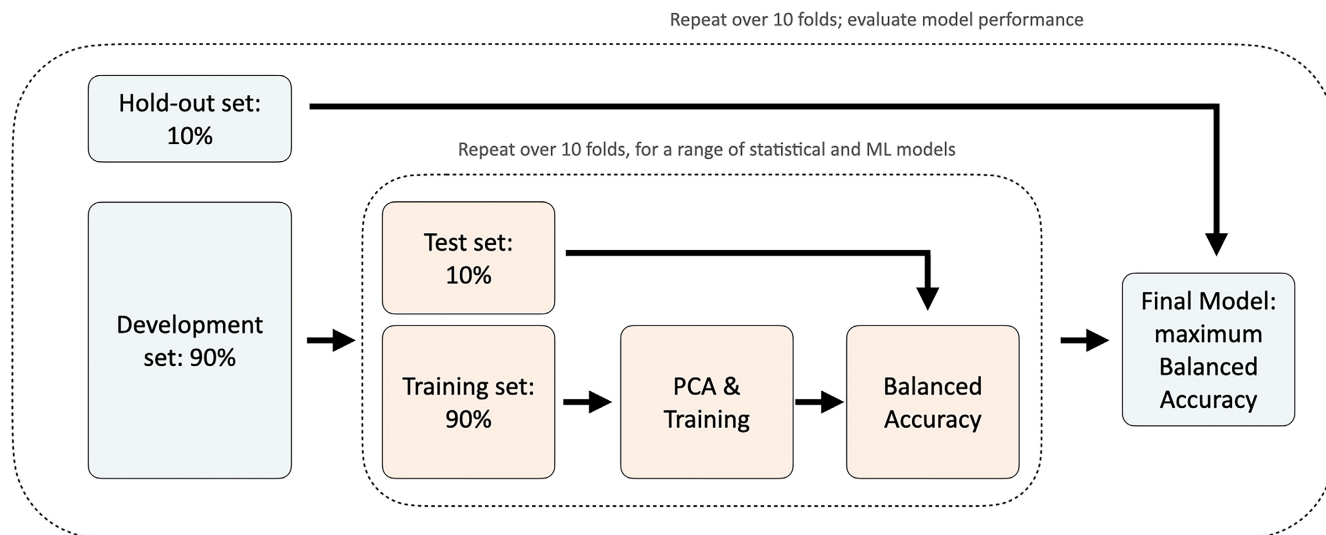


FIGURE 2 Two-tier cross-validation. Flow diagram depicts the two-tier cross-validation scheme for model development. Each individual electroencephalogram is part of a hold-out set exactly once. A suite of statistical models is explored to find a model with optimal performance in terms of balanced accuracy during the training phase. Then this model is tested against the independent hold-out sets. Balanced accuracy is reported for the cross-validated development set and the final hold-out test set. Full details on the modeling steps are given in Appendix S1. PCA, principal component analysis.

2.6 | Evaluating model performance

To evaluate the model performance, we contrasted it with the performances of a naïve classifier, permutation-based models, and a confounder-only model. The naïve classifier corresponds to taking the majority class as the final classification for each subject. In the permutation-based testing, we used 100 permutations without replacement of the labels before cross-validation. This effectively provides a null model, by assessing the maximal performance of the machine learning framework when it is applied to these permuted datasets. This effectively establishes a lower bound for what could be achieved by "chance" alone. Finally, the confounder-only model utilizes the same machine learning framework, but now the features correspond to the confounder data (age, sex, ASM status, and comorbidity status). This effectively provides a baseline for what potential added performance could be derived from the EEG biomarkers.

3 | RESULTS

A total of 814 EEG files from 648 individual patients were considered, having been collected as described in Section 2. All metadata were analyzed by an experienced clinical scientist (L.E.S.). A total of 367 were excluded from the present analysis either because they had no noncontributory EEGs, there were critically missing metadata, or the EEG was provided in an unreadable data format (see also Figure 1). The first recorded noncontributory EEGs

from 281 participants were analyzed for model development and validation. Of these, 89 subjects (31.7%) were reported by the consultant neurophysiologist as nonspecifically abnormal and 192 (68.2%) as normal. Of the 281 subjects, 129 (45.9%) were ultimately diagnosed with epilepsy (nine generalized, 106 focal, and 14 unclassified—either unknown or information not available), and 152 (54.1%) were ultimately diagnosed with an alternative condition. Full details on age, gender, comorbidity, and treatment status are summarized in Table 1.

For the 197 subjects with a normal noncontributory EEG, the random undersampling boosting (RUB) model achieved the highest level of mean balanced accuracy (65.1%, SD = 3.0%) during 10-fold cross-validation in the training phase.²⁷ The PCA approach reduced the feature space to four principal components (mean explained variance = 96.8%, SD = .12%). The final RUB model (with four principal components) achieved balanced accuracy of 67.9% on the independent test set (Figure 3), AUROC of .72, and a diagnostic odds ratio of 4.64 (95% CI = 2.47–8.71). An extensive set of performance metrics, including AUROC, recall, F1 score, Brier score, and a confusion matrix, was calculated (Figure S2); an additional analysis with the mean AUROC as the performance metric during cross-validation similarly identified the RUB model with four principal components as optimal (.69, SD = .024).

For the 93 subjects with an abnormal noncontributory EEGs, the subspace discriminant model achieved the highest level of mean balanced accuracy (58.2%, SD = 4.6%) during 10-fold cross-validation in the training

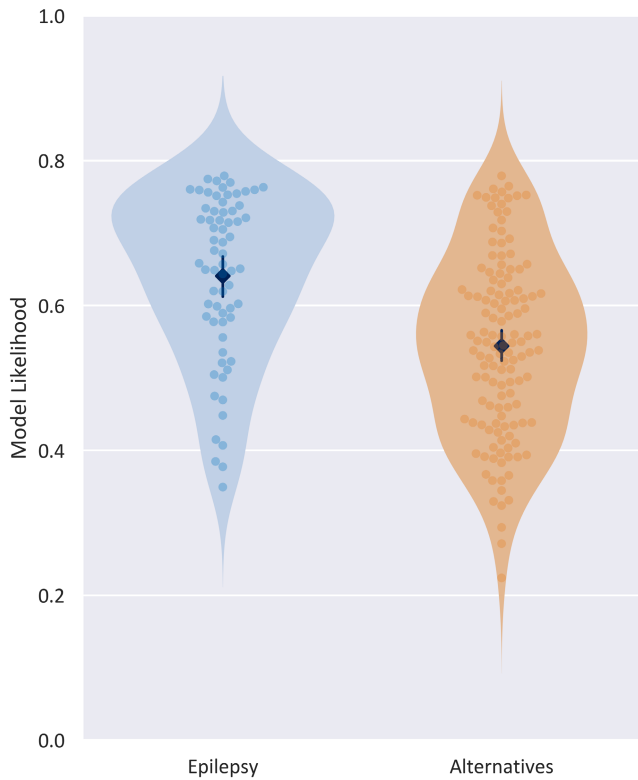


FIGURE 3 Random undersampling boosting (RUB) likelihood scores for normal clinically noncontributory EEGs. Violin plots display the likelihood scores from the RUB model across the epilepsy (blue) and alternative condition (orange) groups for subjects whose EEGs were considered clinically normal but noncontributory. Individual values for each EEG (subject) are displayed with a swarm plot (darker blue and orange). Mean (diamond) and 95% confidence intervals are displayed within each violin plot (black).

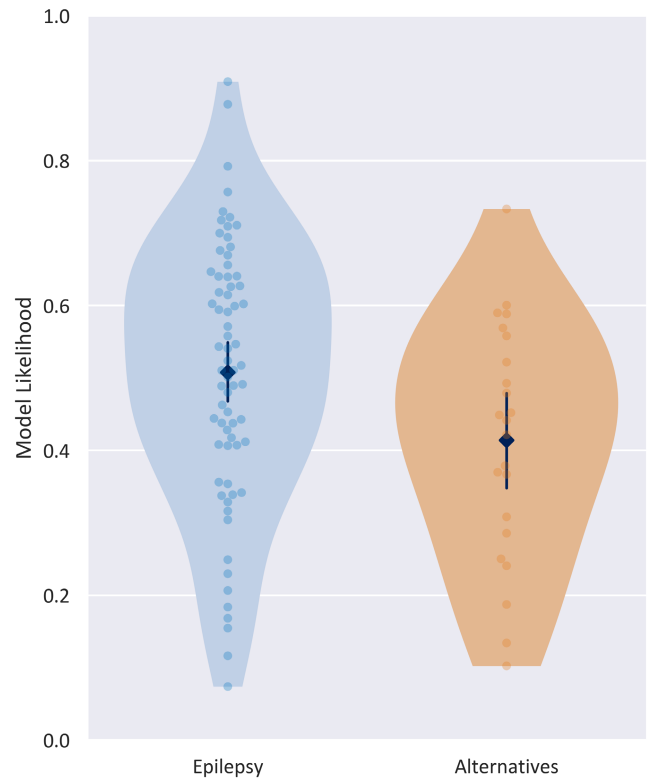


FIGURE 4 Subspace discriminant likelihood scores for abnormal clinically noncontributory electroencephalograms (EEGs). Violin plots display the likelihood scores from the subspace discriminant model across the epilepsy (blue) and alternative condition (orange) groups for subjects whose EEGs were considered clinically abnormal but noncontributory. Individual values for each EEG (subject) are displayed with a swarm plot (darker blue and orange). Mean (diamond) and 95% confidence intervals are displayed within each violin plot (black).

phase. The PCA approach did not reduce the feature space to lower dimensionality. The final subspace discriminant model achieved balanced accuracy of 61.2% on the independent test set (Figure 4), AUROC of .65, and a diagnostic odds ratio of 2.48 (95% CI = .94–6.51; see Figure S2 for an extensive list of performance metrics).

At the level of the individual markers, several of the overall observed trends within the noncontributory EEGs were concordant with the findings from the original development studies using boxplots. In particular, there was an observed trend for increased mean degree, degree variance, and average clustering coefficient, and decreased characteristic path length and critical coupling (see Figures S3–S6). The discordant cases were further investigated. We found no association between misclassification and age, gender, ASM status, the presence of comorbidity, or specific clinical site (Appendix S1).

To assess how model performance contrasts with naïve or chance-driven approaches (e.g., finite sample-size

effects), we compared model performance for the normal noncontributory EEGs against three different independent methods: (1) naïve classification (majority class), (2) permutation-based testing ($n = 100$), and (3) confounder-only model. For the naïve classification, standard accuracy was 54.1% overall (majority: nonepilepsy), 66.0% in the clinically normal cohort and 28.0% in the abnormal cohort due to the inverted imbalances, and 50% throughout when balanced accuracy was the primary outcome metric. For the permutation-based testing, we found a mean balanced accuracy of 50.5% for the normal noncontributory cohort (range = 43.5%–58.8%; see also Figure 5). Finally, the confounder-model reached balanced accuracy of 57.6% for a quadratic support vector machine (SVM) model. When using the AUROC as the primary outcome metric during model development, we found mean AUROC of .50 (range = .40–.60) for the permutation-based tests and .44 for a cubic SVM model using confounders only.

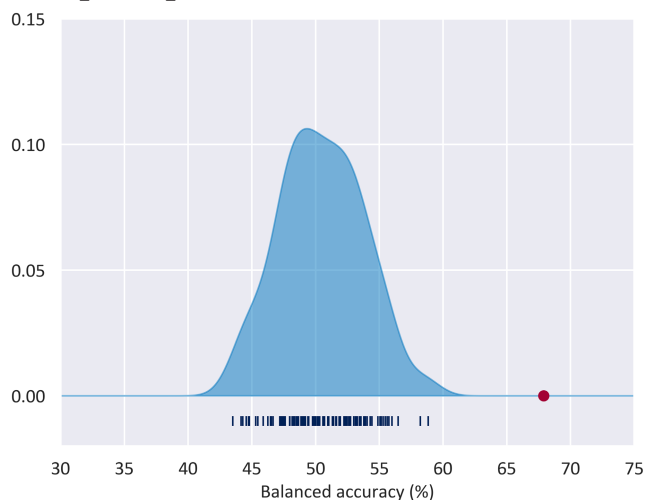


FIGURE 5 Comparing permutation-based models to random undersampling boosting (RUB) performance. Estimated density (using kernel density estimation; Gaussian width=1.2) of 100 permutation-based models assessed by balanced accuracy, with the individual values in a rug plot on the horizontal axis. RUB model performance is displayed as a single red dot.

4 | DISCUSSION

In this retrospective, multisite case–control study, clinically noncontributory EEG recordings were interrogated using a set of candidate biomarkers and a statistical model developed for classification. The models were cross-validated and tested on noncontributory EEGs from a cohort of 129 subjects with epilepsy and 152 subjects originally suspected of having epilepsy but ultimately receiving a diagnosis for an alternative condition. The models achieved increased performance in comparison to existing similar studies, contrasting favorably with the clinical yield for routine EEG recordings of this nature.^{2,7} Additionally, because the classifiers were tested on the first available noncontributory EEG recording for each subject without requiring clinical marking, the developed statistical models could potentially reduce diagnostic delay without increasing clinical workload.

The cohort of people enrolled in the study, who were ultimately diagnosed by a clinical expert either with epilepsy or with an alternative diagnosis, represent the variability of subjects expected in standard clinical practice in a way that single-site studies using healthy participants for control data cannot. Furthermore, in current clinical practice, given that a contributory EEG (e.g., one with the presence of IEDs) can support a diagnosis of epilepsy, additional information gleaned from a digital biomarker is unlikely to alter clinical decision-making or clinical management; in this instance, the presence of IEDs will take precedence. Therefore, the true value in digital biomarkers of whole-brain networks lies in their ability to

discriminate EEGs from persons with and without epilepsy when applied to noncontributory EEGs. Hence, to determine the potential added value of this set of candidate biomarkers, we assessed their performance only in first EEGs that were noncontributory. The routine clinical analysis of these recordings did not contribute one way or the other to a diagnosis at the time they were collected. This is a key differentiator of this study from other studies into candidate biomarkers for epilepsy, which typically include EEGs both containing IEDs and those without, or used high-density EEG.^{5,7,28,29} The approach presented here allows the potential added value of using biomarkers in addition to current standard clinical practice to be assessed. It is important to emphasize that the performance metrics (e.g., sensitivity or balanced accuracy) consider the model as a standalone test (which is not representative of clinical decision-making of epilepsy), rather than as a potential complementary test that could be utilized when the EEG is considered clinically noninformative. In line with this, we emphasize that the presented cross-validated results present *historic* or *retrospective* confidence in the biomarker model. Their ultimate utility within the clinical pathway requires a number of additional clinical validation steps. For example, to assess the added value of clinical decision support, a continuous risk score should be well calibrated, for example, using Platt scaling (see [Figure S7](#)). The results presented suggest that an existing set of candidate biomarkers that was evidenced on primarily phase 1 evidence, contained generalizable predictive power in a retrospective phase 2 setting. It is important to consider how the observed results compare to “null models” (e.g., due to finite sample-size effects); using independent methods (including a confounder-only and permutation-based testing where labels are randomly shuffled prior to model training), we find that the identified model outperforms these methods by a significant margin, providing further confidence in its robustness.

The presence of nonspecific abnormalities led to significantly decreased performance across the performance metrics, whereas other potential confounding variables (age, sex, comorbidity, medication status, clinical site) did not specifically contribute to discordant cases. Interestingly, the overall decrease was primarily caused by decreased specificity, with similar sensitivities, although the modest sample size and class imbalance should be considered (e.g., a naïve majority classifier could achieve similar performance in terms of standard accuracy, but not balanced accuracy). Our results may indicate that the nonspecific abnormalities mediate the average properties of these biomarkers over the entirety of the EEG recording rather than having a pronounced, isolated effect at small sections of the EEGs. This finding suggests that the abnormal model might benefit from including additional

nonlinear, event-driven biomarkers as well as further studies that aim to specifically disambiguate the effect of local abnormalities on background properties. The size of the datasets prohibited further analysis of, for example, the relationship between type of abnormality and impact on the background features, or whether there are certain differential conditions that are more strongly impacted by this.

Given the nature of the considered clinical cohort, this study has several limitations. First, all EEG recordings were collected from subjects aged 18 years and older and therefore may not generalize to pediatric settings. Second, correcting for potential confounders with a basic regression model does not necessarily eliminate their full impact on either EEG features or comparisons. Third, although this is to the best of our knowledge the largest study of its size (i.e., cross-validated, multisite, clinically noncontributory EEG recordings), both statistical models were developed in unbalanced datasets with relatively modest sample sizes, in particular for the smallest class. Various other valid performance metrics could have been used during cross-validation (e.g., AUROC). Furthermore, the confounder model only considered ASM and comorbidities as binary classes rather than multiple classes (e.g., type of comorbidity) or continuous (e.g., drug load). Fourth, the separation between a subject's diagnosis categorization as "epilepsy" or "alternative" does not map directly onto clinical decision-making where the working diagnosis will ultimately be a specific condition or diagnosis (i.e., broad screening rather than a specific final diagnosis). Fifth, some participants had missing data, for example, as a consequence of certain sites being unable to provide information relating to ASM status. In these cases, we chose not to alter scores with the confounder model, because this has the potential to introduce bias. We conducted a post hoc analysis of the performance of the model on a site-by-site basis. We did not find any significant differences between sites (see Error Analysis in Appendix S1). This suggests that missing data are not likely to have a major impact on our findings. Finally, there are a number of sources of potential variation that were not controlled for. These include lack of standardized criteria across clinical sites (both in terms of EEG reading as well as in diagnostic decision-making), influence of recording time, subject alertness levels, and the inclusion of activation methods (e.g., intermittent photic stimulation, hyperventilation). These factors were not explicitly controlled for in this study, to maximize generalizability and applicability of our findings to the clinical setting, where EEG protocols have inherent variability based on local procedures and practice. Nevertheless, these factors may have impacted key features within the EEG.

There are several important steps required to further assess the potential of these types of classification models.

First, the models should be extended to account for the observed decrease in performance in the presence of non-specific abnormalities, by using large datasets and ideally longer recordings to study the dynamic impact between these abnormalities and background features. Second, the models should be tested and verified against a large, independent retrospective dataset from at least one new clinical site. In addition, future studies could explore the difference between the real-world heterogeneity as in the current study and a standardized protocol (expert group examining the EEG and making the diagnostic decision according to a set of standardized criteria). The intrasubject variability of the biomarkers should be explored in a larger subset of participants with subsequent EEG recordings, with a more sophisticated set of statistical models (e.g., to quantify subject-specific natural variation). This would also allow for a more explicit exploration of diagnostic delay as a clinical parameter of interest on a subject-specific basis. Next, the applicability of these markers to multiple classes could be tested and validated. Instead of separating subjects by ultimate diagnosis (epilepsy or alternative condition), these models should test and refine the biomarkers for a specific diagnosis or syndrome, such as NEAD or temporal lobe epilepsy.³⁰ The methodology could also explicitly consider the preference of clinicians in their intended use and context, for example, by prioritizing specificity over sensitivity during the development of the algorithms. Another study should specifically consider the specific variability and variance found across clinical sites and within cohorts, including EEG durations (including longitudinal), activation procedures, comorbidity, and presence of (epileptiform) abnormalities. Finally, it would be particularly interesting to assess their performance at tertiary centers where people with suspected refractory epilepsy are investigated for potential surgery.

We have developed statistical models that may offer improved diagnostic yield for epilepsy in a complementary fashion. This was achieved using a set of digital biomarkers derived from what would currently be considered clinically noncontributory EEG recordings. A promising direction for clinical translation could be the early identification of people with epilepsy after initial noncontributory EEGs for prolonged EEG monitoring, thereby reducing periods of watchful waiting and diagnostic delay. Future prospective and longitudinal studies are crucial to assess the overall utility of these digital markers for the purpose of diagnostic decision support.

AUTHOR CONTRIBUTIONS

Wessel Woldman, John R. Terry, and Rohit Shankar conceived the study and acquired funding. Lydia E. Staniaszek,

Kay Meiklejohn, David Allen, Phil Tittensor, Francesco Manfredonia, Charlotte Lawthom, Matthew C. Walker, Al Anzari Abdul Azeez, Chris Price, Sophie Georgiou, Elizabeth Galizia, David Martin-Lopez, Manny Bagary, and Sakh Khalsa contributed to data acquisition. Luke Tait and Wessel Woldman conducted the data analysis and data visualization. Rohit Shankar and Lydia E. Staniaszek verified the data and reviewed the analysis. Luke Tait, Lydia E. Staniaszek, Rohit Shankar, Benjamin B. Howes, John R. Terry, and Wessel Woldman contributed to data interpretation. Rohit Shankar supervised the study. Luke Tait, Lydia E. Staniaszek, and Wessel Woldman wrote the first draft, and all other authors reviewed and commented on the report. All authors had full access to the complete analysis results of the study and had final responsibility for the decision to submit for publication.

AFFILIATIONS

¹Cardiff University, Cardiff, UK

²University of Birmingham, Birmingham

³University Hospitals Bristol and Weston National Health Service Foundation Trust, Bristol, UK

⁴Neuronostics, Bristol, UK

⁵St. George's Hospital National Health Service Foundation Trust, London, UK

⁶Kingston Hospital National Health Service Foundation Trust, Kingston, UK

⁷University College London, London, UK

⁸University College London Hospitals, London, UK

⁹University Hospital Southampton National Health Service Foundation Trust, Southampton, UK

¹⁰Royal Devon and Exeter National Health Service Foundation Trust, Exeter, UK

¹¹Birmingham and Solihull Mental Health National Health Service Foundation Trust, Birmingham, UK

¹²Royal Wolverhampton National Health Service Trust, Wolverhampton, UK

¹³University of Wolverhampton, Wolverhampton, UK

¹⁴Royal Gwent Hospital, Newport, UK

¹⁵Swansea University, Swansea, UK

¹⁶University of Plymouth, Plymouth, UK

¹⁷Cornwall Partnership National Health Service Foundation Trust, Bodmin, UK

ACKNOWLEDGMENTS

L.T. was supported by the NIHR (AI01646). W.W. was supported by Epilepsy Research UK (F2002), Innovate UK (103939), and the NIHR (AI01646). J.R.T. was supported by the EPSRC (EP/N014391/2 and EP/T027703/1), Innovate UK (103939), and the NIHR (AI01646). This work was supported by the by researchers at the Department of Health's National Institute for Health Research, UCL/UCL Biomedical Research Centre (M.C.W. & A.A.A.A.). R.S. was supported by Innovate UK (103939) and the NIHR (AI01646). We thank Professor Sandor Beniczky for his comments on an earlier draft of the manuscript.

FUNDING INFORMATION

Funding was provided by Innovate UK, National Institute for Health and Care Research, Engineering and Physical Sciences Research Council, and Epilepsy Research UK.

CONFLICT OF INTEREST STATEMENT

L.E.S., K.M., are B.B.H. are employees of Neuronostics. W.W. and J.R.T. are cofounders, directors, and shareholders of Neuronostics. None of the other 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.

DATA AVAILABILITY STATEMENT

The EEG recordings and metadata are not publicly available due to restrictions by privacy laws. Postprocessed data supporting the findings of this study are available upon reasonable request from the corresponding author (W.W.), and additional metadata may be made available upon reasonable request from the study sponsor (R.S.).

ETHICS STATEMENT

This study was approved by the HRA & HCRW (Integrated Research Application System (IRAS)24: 260729) and by the internal ethics committees of all collaborating clinical institutes.

ORCID

Matthew C. Walker  <https://orcid.org/0000-0002-0812-0352>

Rohit Shankar  <https://orcid.org/0000-0002-1183-6933>

Wessel Woldman  <https://orcid.org/0000-0003-2957-6276>

REFERENCES

1. WHO epilepsy factsheet. [cited 2023 Feb 2]. Available from: <http://www.who.int/mediacentre/factsheets/fs999/en/>
2. Smith SJM. EEG in the diagnosis, classification, and management of patients with epilepsy. *J Neurol Neurosurg Psychiatry*. 2005;76:ii2-ii7.
3. Bouma HK, Labos C, Gore GC, Wolfson C, Keezer MR. The diagnostic accuracy of routine electroencephalography after a first unprovoked seizure. *Eur J Neurol*. 2016;23:455-63.
4. NICE guidelines 2017. [cited 2023 Feb 2]. Available from: <https://www.nice.org.uk/guidance/ng217>
5. Faiman I, Smith S, Hodsoll J, Young AH, Shotbolt P. Resting-state EEG for the diagnosis of idiopathic epilepsy and psychogenic nonepileptic seizures: a systematic review. *Epilepsy Behav*. 2021;121:108047.
6. Dharan AL, Bowden SC, Lai A, Peterson ADH, Cheung MWL, Woldman W, et al. Resting-state functional connectivity in the idiopathic generalized epilepsies: a systematic review and meta-analysis of EEG and MEG studies. *Epilepsy Behav*. 2021;124:108336.

7. Thomas J, Thangavel P, Peh WY, Jing J, Yuvaraj R, Cash SS, et al. Automated adult epilepsy diagnostic tool based on interictal scalp electroencephalogram characteristics: a six-center study. *Int J Neural Syst*. 2021;31:1–17.
8. Tveit J, Aurlien H, Plis S, Calhoun VD, Tatum WO, Schomer DL, et al. Automated interpretation of clinical electroencephalograms using artificial intelligence. *JAMA Neurol*. 2023;80:805–12.
9. Nascimento FA, Barfuss JD, Jaffe A, Brandon Westover M, Jing J. A quantitative approach to evaluating interictal epileptiform discharges based on interpretable quantitative criteria. *Clin Neurophysiol*. 2023;146:10–7.
10. van Diessen E, Zweiphenning WJEM, Jansen FE, Stam CJ, Braun KPJ, Otte WM. Brain network organization in focal epilepsy: a systematic review and meta-analysis. *PLoS One*. 2014;9:1–21.
11. Douw L, de Groot M, van Dellen E, Heimans JJ, Ronner HE, Stam CJ, et al. “Functional connectivity” is a sensitive predictor of epilepsy diagnosis after the first seizure. *PLoS One*. 2010;5:1–7.
12. Larsson G, Kostov H. Lower frequency variability in the alpha activity in EEG among patients with epilepsy. *Clin Neurophysiol*. 2005;116:2701–6.
13. Schmidt H, Woldman W, Goodfellow M, Chowdhury FA, Koutroumanidis M, Jewell S, et al. A computational biomarker of idiopathic generalized epilepsy from resting state EEG. *Epilepsia*. 2016;57:e200–e204.
14. Oostenveld R, Fries P, Maris E, Schoffelen JM. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci*. 2011;2011:1–9.
15. Schmidt H, Petkov G, Richardson MP, Terry JR. Dynamics on networks: the role of local dynamics and global networks on the emergence of hypersynchronous neural activity. *PLoS Comput Biol*. 2014;10:e1003947.
16. Soriano MC, Niso G, Clements J, Ortín S, Carrasco S, Gudín M, et al. Automated detection of epileptic biomarkers in resting-state interictal MEG data. *Front Neuroinform*. 2017;11:1–12.
17. Lopes MA, Krzemiński D, Hamandi K, Singh KD, Masuda N, Terry JR, et al. A computational biomarker of juvenile myoclonic epilepsy from resting-state MEG. *Clin Neurophysiol*. 2021;132:922–7.
18. Varatharajah Y, Berry B, Joseph B, Balzekas I, Pal Attia T, Kremen V, et al. Characterizing the electrophysiological abnormalities in visually reviewed normal EEGs of drug-resistant focal epilepsy patients. *Brain Commun*. 2021;3:1–17.
19. Wadhera T. Brain network topology unraveling epilepsy and ASD association: automated EEG-based diagnostic model. *Expert Syst Appl*. 2021;186:115762.
20. Shakeshaft A, Laiou P, Abela E, Stavropoulos I, Richardson MP, Pal DK, et al. Heterogeneity of resting-state EEG features in juvenile myoclonic epilepsy and controls. *Brain Commun*. 2022;4:fcac180.
21. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:1–12.
22. Yaakub SN, Tangwiriyasakul C, Abela E, Koutroumanidis M, Elwes RDC, Barker GJ, et al. Heritability of alpha and sensorimotor network changes in temporal lobe epilepsy. *Ann Clin Transl Neurol*. 2020;7:667–76.
23. Pegg EJ, Taylor JR, Mohanraj R. Spectral power of interictal EEG in the diagnosis and prognosis of idiopathic generalized epilepsies. *Epilepsy Behav*. 2020;112:107427.
24. Shackman AJ, McMenamin BW, Maxwell JS, Greischar LL, Davidson RJ. Identifying robust and sensitive frequency bands for interrogating neural oscillations. *Neuroimage*. 2010;51:1319–33.
25. Chowdhury FA, Woldman W, FitzGerald TH, Elwes RD, Nashef L, Terry JR, et al. Revealing a brain network endophenotype in families with idiopathic generalised epilepsy. *PLoS One*. 2014;9:e110136.
26. Rubinov M, Sporns O. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*. 2010;52:1059–69.
27. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: improving classification performance when training data is skewed. *Proc Int Conf Pattern Recognit*. 2008;1–4. <https://doi.org/10.1109/ICPR.2008.4761297>
28. Coito A, Genetti M, Pittau F, Iannotti GR, Thomschewski A, Höller Y, et al. Altered directed functional connectivity in temporal lobe epilepsy in the absence of interictal spikes: a high density EEG study. *Epilepsia*. 2016;57:402–11.
29. Verhoeven T, Coito A, Plomp G, Thomschewski A, Pittau F, Trinka E, et al. Automated diagnosis of temporal lobe epilepsy in the absence of interictal spikes. *Neuroimage Clin*. 2018;17:10–5.
30. Woldman W, Schmidt H, Abela E, Chowdhury FA, Pawley AD, Jewell S, et al. Dynamic network properties of the interictal brain determine whether seizures appear focal or generalised. *Sci Rep*. 2020;10:1–11.

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: Tait L, Staniaszek LE, Galizia E, Martin-Lopez D, Walker MC, Azeez AAA, et al. Estimating the likelihood of epilepsy from clinically noncontributory electroencephalograms using computational analysis: A retrospective, multisite case-control study. *Epilepsia*. 2024;00:1–11. <https://doi.org/10.1111/epi.18024>