Received: 27 November 2020 | Revised: 25 February 2021 | Accepted: 27 February 2021

DOI: 10.1111/ane.13418

ORIGINAL ARTICLE

Acta Neurologica Scandinavica WILEY

Early identification of epilepsy surgery candidates: A multicenter, machine learning study

Benjamin D. Wissel[1] | Hansel M. Greiner[2,3] | Tracy A. Glauser[2,3] | John P. Pestian[1,2] | Andrew J. Kemme[4] | Daniel Santel[1] | David M. Ficker[5] | Francesco T. Mangano[2,6] | Rhonda D. Szczesniak[2,7] | Judith W. Dexheimer[1,2,4]

[1]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

[2]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

[3]Division of Neurology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

[4]Division of Emergency Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

[5]Department of Neurology and Rehabilitation Medicine, University of Cincinnati, Cincinnati, OH, USA

[6]Division of Neurosurgery, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

[7]Division of Biostatistics & Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Correspondence

Judith Dexheimer, Department of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, MLC 2008, 3333 Burnet Avenue, Cincinnati, OH 45229-3039, USA.
Email: judith.dexheimer@cchmc.org

Funding information

National Institute of Neurological Disorders and Stroke, Grant/Award Number: F31 NS115447; Agency for Healthcare Research and Quality, Grant/Award Number: R21 HS024977; University of Cincinnati's Center for Health Informatics

Objectives: Epilepsy surgery is underutilized. Automating the identification of potential surgical candidates may facilitate earlier intervention. Our objective was to develop site-specific machine learning (ML) algorithms to identify candidates before they undergo surgery.

Materials & Methods: In this multicenter, retrospective, longitudinal cohort study, ML algorithms were trained on n-grams extracted from free-text neurology notes, EEG and MRI reports, visit codes, medications, procedures, laboratories, and demographic information. Site-specific algorithms were developed at two epilepsy centers: one pediatric and one adult. Cases were defined as patients who underwent resective epilepsy surgery, and controls were patients with epilepsy with no history of surgery. The output of the ML algorithms was the estimated likelihood of candidacy for resective epilepsy surgery. Model performance was assessed using 10-fold cross-validation.

Results: There were 5880 children (n = 137 had surgery [2.3%]) and 7604 adults with epilepsy (n = 56 had surgery [0.7%]) included in the study. Pediatric surgical patients could be identified 2.0 years (range: 0–8.6 years) before beginning their presurgical evaluation with AUC =0.76 (95% CI: 0.70–0.82) and PR-AUC =0.13 (95% CI: 0.07–0.18). Adult surgical patients could be identified 1.0 year (range: 0–5.4 years) before beginning their presurgical evaluation with AUC =0.85 (95% CI: 0.78–0.93) and PR-AUC =0.31 (95% CI: 0.14–0.48). By the time patients began their presurgical evaluation, the ML algorithms identified pediatric and adult surgical patients with AUC =0.93 and 0.95, respectively. The mean squared error of the predicted probability of surgical candidacy (Brier scores) was 0.018 in pediatrics and 0.006 in adults.

Conclusions: Site-specific machine learning algorithms can identify candidates for epilepsy surgery early in the disease course in diverse practice settings.

KEYWORDS

artificial intelligence, electronic health record, epilepsy, machine learning, medical informatics, neurology

1 | INTRODUCTION

There are more than 45 million people with epilepsy.[1] One-third of patients do not adequately respond to anti-epileptic drugs (AEDs).[2] Undergoing resective neurosurgery increases the chances of seizure freedom in eligible patients from 10% to 67%.[3–5] Yet surgery is highly underutilized,[3,6] resulting in decreased quality of life,[7] increased mortality,[8] poorer cognitive outcomes,[9] and increased healthcare costs.[10] Surgical referrals can be delayed due to a lack of standardized referral processes and a perceived lack of resources.[11] Automating the identification of surgical candidates can help to reduce the number of eligible patients that are missed.

Machine learning methodologies (ML) can be used to identify candidates for epilepsy surgery years before they undergo surgery.[12–15] ML algorithms are infrequently implemented into care. In pediatrics, one algorithm was fully automated including the provision of decision support to providers[13]; however, the positive predictive value (PPV) of the alerts from this system was low at only 25%.[13] Improving the accuracy of this algorithm will increase its impact on patient care. One drawback of the current algorithm is that it is limited to neurology visit notes. It does not use other relevant information from the electronic health record (EHR).

Additionally, ML applications have never been used to identify surgical candidacy in the adult population. The principles of selecting children and adults with epilepsy for a presurgical evaluation are similar, but key differences exist.[3,16] Therefore, it would be difficult to design an algorithm that could be trained to work well in both settings.[17] A methodology to account for the heterogeneity in epilepsy phenotypes and care patterns is needed.

We sought to create a ML methodology that can be deployed at epilepsy centers with disparate patient populations and care patterns (e.g., primary vs. tertiary, children vs. adults, etc.). The purpose of this study was to develop a generalizable methodology that could be used to build site-specific ML algorithms that identify candidates for epilepsy surgery before they undergo surgery.

2 | METHODS

2.1 | Standard protocol approvals, registrations, and patient consents

This study was approved by the Cincinnati Children's Hospital Medical Center (CCHMC) institutional review board (approval #2012-1646 and #2016-1932) with a waiver for informed consent. The approval applied to both hospitals included in this study. This report adheres to the Transparent Reporting of studies on prediction models for Individual Prognosis Or Diagnosis (TRIPOD) guidelines.[18]

2.2 | Study population

We conducted a retrospective analysis of EHR data from two healthcare systems, CCHMC and the University of Cincinnati Medical Center (UC). The hospitals at CCHMC and UC are distinct institutions with their own EHR installations.

CCHMC is comprised of two pediatric hospitals and 14 outpatient clinic sites and is centrally located in a metropolitan area with over 2 million people. There are approximately 6000 unique patients with epilepsy who visit CCHMC annually. All patients who visited a CCHMC neurology clinic for a diagnosis of epilepsy between 23 January 2009 and 4 April 2019 were eligible for inclusion. Epilepsy diagnoses were made clinically and defined here according to International Classification of Diseases (ICD) codes (G40.* or the ICD-9 equivalents), which are sensitive and have a high positive predictive value for epilepsy.[19] To retain patients who established neurology care, patients with less than two neurology visits were excluded.

UC is comprised of two main campuses and 27 outpatient clinic sites. There are 4000 unique patients with epilepsy who visit UC annually. All patients who visited at UC neurology clinic for a diagnosis of epilepsy between 10 July 2012 and 10 May 2019 were eligible for inclusion, and similar exclusion criteria were applied. Start dates corresponded to the dates that the EHR was operationalized at the respective institutions.

Eligible patients were assigned into two groups: (1) non-surgical and (2) history of resective surgery. Surgical status was defined using Current Procedural Terminology (CPT) codes and confirmed by manual chart review. Resective surgical procedures included lobectomies, corticectomies, lesionectomies, multi-lobar surgeries, and hemispherectomies. Similar to past work,[13] patients with a history of palliative surgery, such as the implantation of a vagus nerve stimulator, responsive neurostimulation, deep brain stimulation, and corpus collosotomy, were excluded since the phenotype of patients who are eligible for palliative surgeries is different than those potentially eligible for a resective surgery.

The ML algorithms were trained to predict whether a patient was in the non-surgical or resective surgery group. Outputs from each model were compared to clinical outcomes: resective surgery vs. no surgery. Study personnel manually reviewed the charts of every surgical patient to determine the date that they entered the presurgical evaluation protocol. Data generated after patients began their evaluation were discarded to prevent label leak. Only data from before patients entered the presurgical evaluation protocol were used. We excluded surgical patients who were referred to these centers specifically to be evaluated for surgery.

2.3 | Experimental setup and model evaluations

The modeling process, defined here as data extraction, preprocessing, feature extraction, feature selection, and model training, was developed at the pediatric (CCHMC) center. Once established, the modeling process was validated at the adult (UC) center. The procedures for building the adult model were the same as the pediatric model, but the resulting feature sets and algorithms used to classify pediatric and adult patients were different. This allowed models to account for the heterogeneity in the patient populations and care patterns at the different centers. All feature selection was performed within each cross-validation fold.

2.4 | Neurology notes

Neurology notes were processed according to the existing state-of-the-art NLP methods in this domain.[13] We extracted visit notes written by an attending physician, fellow physician, resident physician, or nurse practitioner that were at least 100 words in length. Notes for each patient were concatenated. Years in the text were replaced with the string "_YEAR" and all other numerals with "_NUM". Generic, trade name, and abbreviations for AEDs were mapped to unifying medication codes. For example, "Depakote", "valproic acid", and "VPA" were all mapped to "DRUG_VPA". The *quanteda*[20] package was used to remove all capitalization, punctuation, symbols, separators, and stop words. Free text in the notes was tokenized unigram features. Unigram frequencies were normalized to Boolean values to indicate their presence or absence in the notes. This minimized the effect of copy forwarding. Unigrams present in fewer than ten patients' notes in the pediatric dataset and 15 patients' notes in the adult dataset were removed.

2.5 | EEG and MRI reports

We identified EEGs and head MRIs and extracted the corresponding free-text radiology reports containing the radiologist's written summary and interpretation. We concatenated each patients' EEG and MRI reports under the assumption that features present in MRI reports would have similar medical significance if they appeared in an EEG report. Text pre-processing followed a similar protocol as neurology notes, except that unigrams present in at least two patients' notes were included.

2.6 | Structured data

Features for structured data were chosen based on literature review[16,21–23] and expert opinion (H.M.G., T.A.G., D.M.F, and F.T.M.). We extracted demographics, insurance, outpatient and emergency room visits, hospitalizations, orders for AEDs, procedures, and laboratory values. Demographic information was represented by categorical variables for age, gender, race, insurance type, and distance from the pediatric or adult hospitals. Cutoffs were chosen based on prior literature.[15,24] Zip codes were used to determine how far patients traveled to receive care. Median household incomes for zip codes were obtained from the United States Census Bureau[25] and represented as a continuous variable. Mean imputation was used when there was no income data for a patient's zip code. We represented prescriptions for AEDs as Boolean features and tabulated the total number of AEDs and number of new-generation AEDs (lamotrigine, levetiracetam, oxcarbazepine, and zonisamide)[26] for each patient. To capture epilepsy disease burden, hospitalizations and neurology, neurosurgery, social work, and emergency room visits were each represented as three continuous variables: total number of visits, visits per year, and the greatest number of visits within any six-month period. Hospitalizations for any reason and hospitalizations with epilepsy or seizures listed as the primary reason for the encounter were counted separately. The same was done for emergency room visits. Duration of follow-up was defined as the number of days between each patient's earliest and most recent neurology visits.

Procedures and laboratory orders were represented as Boolean features. This resulted in sparse feature vectors, so we discarded procedures and laboratories that were performed in <2% of surgical patients and <2% of non-surgical patients. No other feature selection was performed for structured data.

2.7 | Feature selection

To reduce the dimensionality of the EHR-based feature vectors for each patient, we used a correlation-based filter to select the 100 most important unigram features from neurology notes and EEG and MRI reports. Unigrams were ranked using a two-sample Kolmogorov-Smirnov test within each cross-validation fold, and the number of features to include was empirically selected from cross-validation results. We used random forest variable importance to select the 100 most importance features from structured data. Models were built using the top 20, 50, and 300 features from the three data modalities. We used an unbiased impurity-based variable importance measure[27] from the *ranger* package[28] in R to select the most important features. Features associated with the presurgical evaluation process were excluded (see list in Table S1). Feature selection was performed within each cross-validation fold.

We compared the performance of several ML algorithms, including logistic regression, L1-regularized logistic regression (least absolute shrinkage and selection operator [LASSO]), L2-regularized logistic regression, support vector machine, gradient boosted machine, random forest, and multilayer perceptrons (MLP) with one, two, and three hidden layers. A random forest with 10 000 trees performed the best and was included. The random forest was built using the *ranger* package with default settings.[28]

2.8 | Interpreting the most important features in the electronic health record

Variable importance plots were used to display the 50 most important features in the EHR. A feature selection procedure similar to the early fusion model was used, except that feature selection was performed using all patients in each dataset. These features were not used for prediction, as they were selected outside the cross-validation folds.

2.9 | Comparison to baseline models

The models were compared to two baseline logistic regression models that predicted surgical status using (1) baseline demographic variables, including age, sex, and race, and duration of follow-up or (2) baseline demographics and duration of follow-up plus the number AEDs prescribed.

2.10 | Early identification

To evaluate whether ML can be used to identify surgical patients early in the disease course, the models were applied, unaltered, to patient data that was censored at the date of their second neurology visit. Surgical candidacy scores at patients' second visit were compared to their surgical status later in the disease course.

2.11 | Statistical analysis

A stratified 10-fold cross-validation scheme was used to assess the internal performance of each model. The primary performance metric was area under the receiver operating characteristic (ROC) curve (AUROC). The secondary performance metric was area under the precision-recall curve (PR-AUC). PR-AUC provides a summary of the precision-recall curve and is a highly discriminant performance marker for imbalanced datasets where the prevalence of the outcome is relatively low.[29] Point estimates and 95% confidence intervals (CI) were estimated by calculating the mean and standard deviation performance within each fold. Model PR-AUCs and AUROCs were compared using t tests. ROC and PR curves were generated using the *precrec* package.[30] To evaluate calibration, we plotted nonparametric calibration curves using the *loess* function with default parameters[31] and quantified the degree to which calibration curves deviated from the identity line, $y = x$. We also calculated the Brier score,[32] which can be interpreted as the mean squared error between the model's estimated likelihood of surgical candidacy and their actual clinical status, which was represented as zero for non-surgical and one for surgical patients. Smaller Brier scores indicate better performance. Two-sided p values <0.05 were considered significant. All analyses were conducted in R statistical software version 3.6.1.[33]

2.12 | Data availability statement

EHR data used in this study cannot be shared because it contains identifiable information protected under the Health Insurance Portability and Accountability Act (HIPAA).

3 | RESULTS

3.1 | Study cohort

There were 5743 non-surgical (51% male; 13.3 ± 7.50 years old) and 137 surgical patients (61% male; 9.74 ± 5.86 years old) included in the pediatric dataset, and 7548 non-surgical (44% male; 47.6 ± 16.8 years old) and 56 surgical patients (50% male; 41.6 ± 12.5 years old) included in the adult dataset (Table 1). The pediatric patients had 6.51 ± 4.82 neurology notes, and the adults had 5.70 ± 4.32 neurology notes included. An EEG or MRI report was available for analysis in 4580 (77.9%) of the pediatric patients and 4313 (56.7%) of the adult patients.

3.2 | Model construction and feature selection

Neurology notes from the pediatric and adult centers were tokenized into 39 100 and 35 838 unigram features, respectively. EEG and MRI reports were tokenized into 8419 and 11 838 unigram features.

For structured data, there were 21 features for demographics, 32 for medications, and 27 for visits and hospitalizations. The pediatric center had 220 procedures and laboratories included, and the adult center had 293 procedures and labs. Income was imputed for 40 (0.7%) pediatric and 22 (0.3%) adult patients. No other demographic variables had missing data.

3.3 | Model performance

Model performance is summarized in Table 2, and the corresponding ROC and PR curves are shown in Figure 1. Baseline models had fair-to-good discrimination in both datasets (AUROC = 0.631–0.852), but their PR-AUC was low (PR-AUC = 0.045–0.194). Including AEDs in the baseline model increased performance in the pediatric dataset (AUROC = 0.852 with AEDs vs. AUROC = 0.631 without AEDs; p < 0.001) but not in the adult dataset (AUROC = 0.740 with AEDs vs. AUROC = 0.739 without AEDs; p = 0.64).

Using information available up to the start of the first presurgical evaluation visit, the random forest model with 50 features performed the best in both the pediatric (AUROC = 0.927 [95% CI: 0.905–0.949]; PR-AUC = 0.417 [95% CI: 0.318–0.517]) and adult datasets (AUROC = 0.946 [95% CI: 0.916–0.976]; PR-AUC = 0.466 [95% CI: 0.289–0.644]).

Using information available at the second visit, pediatric surgical patients could be distinguished from non-surgical patients with AUROC = 0.762 (95% CI: 0.703–0.821), even though surgical patients did not begin their surgical evaluation until 2.0 years later, on average (range: 0–8.6 years). Adult surgical patients could be identified at the second visit, or 1.0 year before their presurgical evaluation (range: 0–5.4 years), with AUROC = 0.851 (95% CI: 0.777–0.926). PR-AUC at the second visit was 0.125 (95% CI: 0.071–0.178) in the pediatric cohort and 0.306 (95% CI: 0.136–0.476) in the adult cohort.

TABLE 1 Patient demographics for each dataset

	Pediatric health system		Adult health system	
Variable	**Non-surgical (*n* = 5743)**	**Surgery (*n* = 137)**	**Non-surgical (*n* = 7548)**	**Surgery (*n* = 56)**
Age, years	13.3 ± 7.50	9.74 ± 5.86	47.6 ± 16.8	41.6 ± 12.5
Male gender	2945 (51.3%)	83 (60.6%)	3340 (44.3%)	28 (50.0%)
Race				
White	4614 (80.3%)	108 (78.8%)	5929 (78.6%)	51 (91.1%)
Black	659 (11.5%)	12 (8.76%)	1363 (18.1%)	5 (8.93%)
Asian	71 (1.24%)	7 (5.11%)	65 (0.86%)	0 (0%)
Other	163 (2.84%)	5 (3.65%)	132 (1.75%)	0 (0%)
Multi-racial	187 (3.26%)	4 (2.92%)	53 (0.70%)	0 (0%)
Unknown	49 (0.85%)	1 (0.73%)	6 (0.08%)	0 (0%)
Insurance[a]				
Private	3064 (53.4%)	63 (46.0%)	3045 (40.3%)	29 (51.8%)
Public	3294 (57.4%)	95 (69.3%)	4262 (56.5%)	25 (44.6%)
Other	55 (0.96%)	3 (2.19%)	241 (3.19%)	2 (3.57%)
Distance from care				
0–25 miles	2727 (47.5%)	61 (44.5%)	5387 (71.4%)	39 (69.6%)
25–50 miles	1006 (17.5%)	24 (17.5%)	1375 (18.2%)	7 (12.5%)
50–100 miles	1004 (17.5%)	30 (21.9%)	586 (7.76%)	8 (14.3%)
>100 miles	1006 (17.5%)	22 (16.1%)	200 (2.65%)	2 (3.57%)
Income, $	54 114 ± 17 691	54 753 ± 16 811	54 667 ± 17 575	52 444 ± 14 839
Number of neurology visits	6.32 ± 4.73	7.93 ± 6.04	6.37 ± 5.04	4.21 ± 4.52
Duration of follow-up, years	2.99 ± 2.58	2.11 ± 2.01	3.04 ± 2.16	1.33 ± 1.69
Anti-epileptic drugs	1.96 ± 1.52	4.12 ± 2.1	2.09 ± 1.45	1.93 ± 1.56
Procedures and labs	14.0 ± 11.1	21.0 ± 15.2	23.3 ± 27.7	12.9 ± 16.3
EEG present	4046 (70.5%)	98 (71.5%)	3336 (44.2%)	22 (39.3%)
MRI present	3026 (52.7%)	102 (74.5%)	3093 (41.0%)	18 (32.1%)

Note: Data are presented as mean ± standard deviation or the number and percent of patients in each group.

Abbreviations: CCHMC, Cincinnati Children's Hospital Medical Center; CI, confidence interval; EEG, electroencephalogram; MRI, magnetic resonance imaging; SD, standard deviation; UC, University of Cincinnati.

[a]Percentages do not add up to 100% because some patients had more than one type of insurance.

3.4 | Important features

Figure 2 shows the top 50 variables variables from neurology notes, EEG and MRI reports, and structured data for the pediatric and adult datasets. The pattern of neurology office visits (e.g., total number, frequency, and duration of follow-up), number and types of AEDs prescribed, and descriptions of seizure types (e.g., “localization” and “mesial”) and drug resistance (e.g., “fails” and “refractory”) were highly informative. Orders for Lacosamide, oxcarbazepine, and carbamazepine, medications commonly used to treat partial epilepsy, were predictive of surgical candidacy in both patient populations.

3.5 | Calibration

Higher surgical candidacy scores were associated with higher likelihood of undergoing surgery (Figure 3). Scores above 0.25, which were assigned to the top 1.5% of children and 0.3% of adults, underestimated the observed probability of surgical candidacy. In adults, scores below 0.25 overestimated the observed probability of surgical candidacy. The mean squared error of the surgical candidacy scores (Brier score) was 0.018 for the pediatric dataset and 0.006 for the adult dataset.

4 | DISCUSSION

Epilepsy is a heterogeneous neurological disorder that has many etiologies and phenotypes,[2] and each epilepsy center has its own unique patient population and clinical workflows. Thus, a ML algorithm trained using EHR data from one center will likely not perform as well at outside centers.[17] The lack of harmonization between EHR systems at different institutions creates an additional barrier.[34] Our solution was to develop a generalizable modeling process

TABLE 2 Model performance in each dataset. The two baseline logistic regression models were labeled as "Demographics + Time" and "Demographics + Time + AEDs," depending on whether anti-epileptic drug prescriptions were included

	Pediatric data		Adult data	
Model	AUROC (95% CI)	PR-AUC (95% CI)	AUROC (95% CI)	PR-AUC (95% CI)
Baseline				
Demographics + time	0.631 (0.591–0.671)	0.045 (0.033–0.056)	0.739 (0.659–0.819)	0.091 (0.038–0.144)
Demographics + time + AEDs	0.852 (0.819–0.885)	0.194 (0.141–0.248)	0.740 (0.660–0.819)	0.069 (0.025–0.112)
Electronic health record-wide data				
Top 20 features	0.915 (0.890–0.939)	0.385 (0.290–0.480)	0.922 (0.876–0.968)	0.415 (0.252–0.577)
Top 50 features	0.927 (0.905–0.949)	0.417 (0.318–0.517)	0.946 (0.916–0.976)	0.466 (0.289–0.644)
Top 100 features	0.929 (0.905–0.952)	0.394 (0.280–0.507)	0.951 (0.929–0.973)	0.471 (0.302–0.640)
Top 300 features	0.924 (0.898–0.950)	0.360 (0.251–0.469)	0.947 (0.924–0.969)	0.480 (0.316–0.644)

Note: The "Electronic Health Record-Wide" algorithms were random forests. Features were selected within each cross-validation fold.

Abbreviations: AED, anti-epileptic drugs; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; PR-AUC, area under the precision-recall curve.

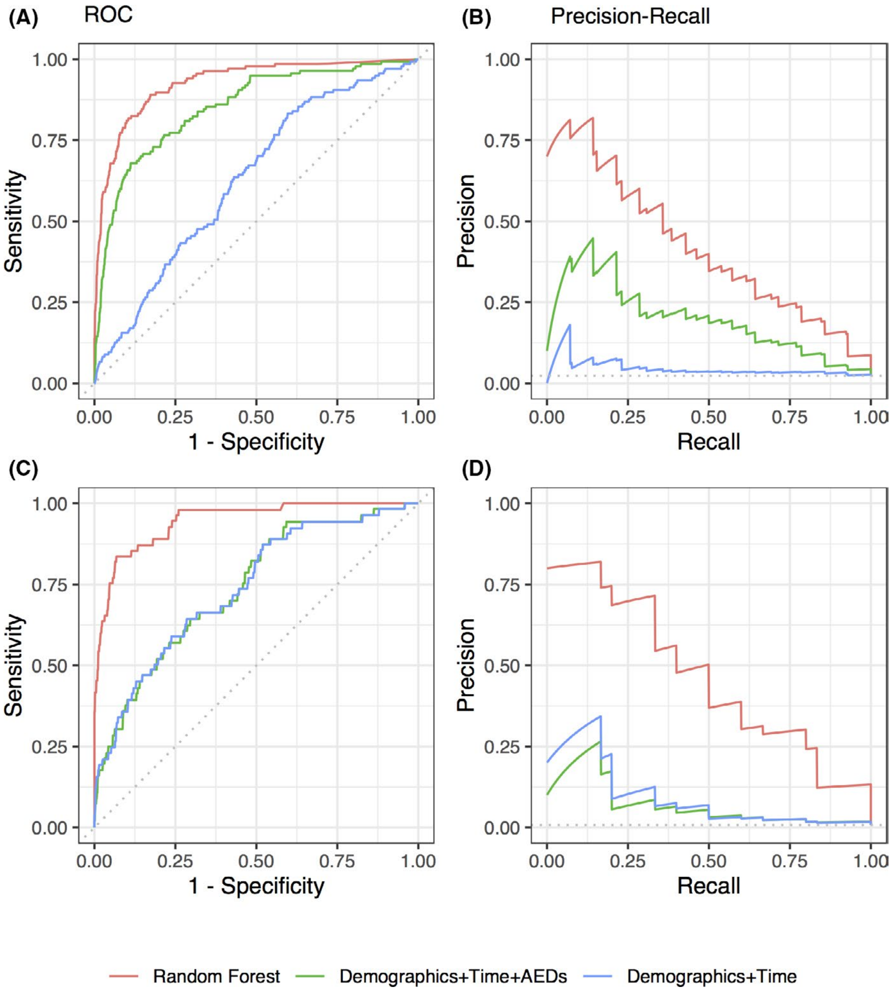

FIGURE 1 Receiver operating characteristic (ROC) and precision-recall curves for the pediatric (A, B) and adult (C, D) datasets. The gray dotted line represents the performance of a random classifier. "Demographics + Time + AEDs" represents the baseline logistic regression model that included anti-epileptic drug prescriptions, and "Demographics + Time" represents the baseline model without anti-epileptic drugs

that encompasses the full data extraction, pre-processing, feature extraction and selection, and model training procedure. This enabled two epilepsy centers, one adult and one pediatric, to create site-specific algorithms that identified surgical candidates in their patient population. Performance was strong at both sites as surgical patients could be identified as early as their second visit, well before most patients began their presurgical evaluation.

There were important differences between the pediatric and adult datasets that were reflected in the results. Pediatric patients typically present earlier in the disease course, sometimes after

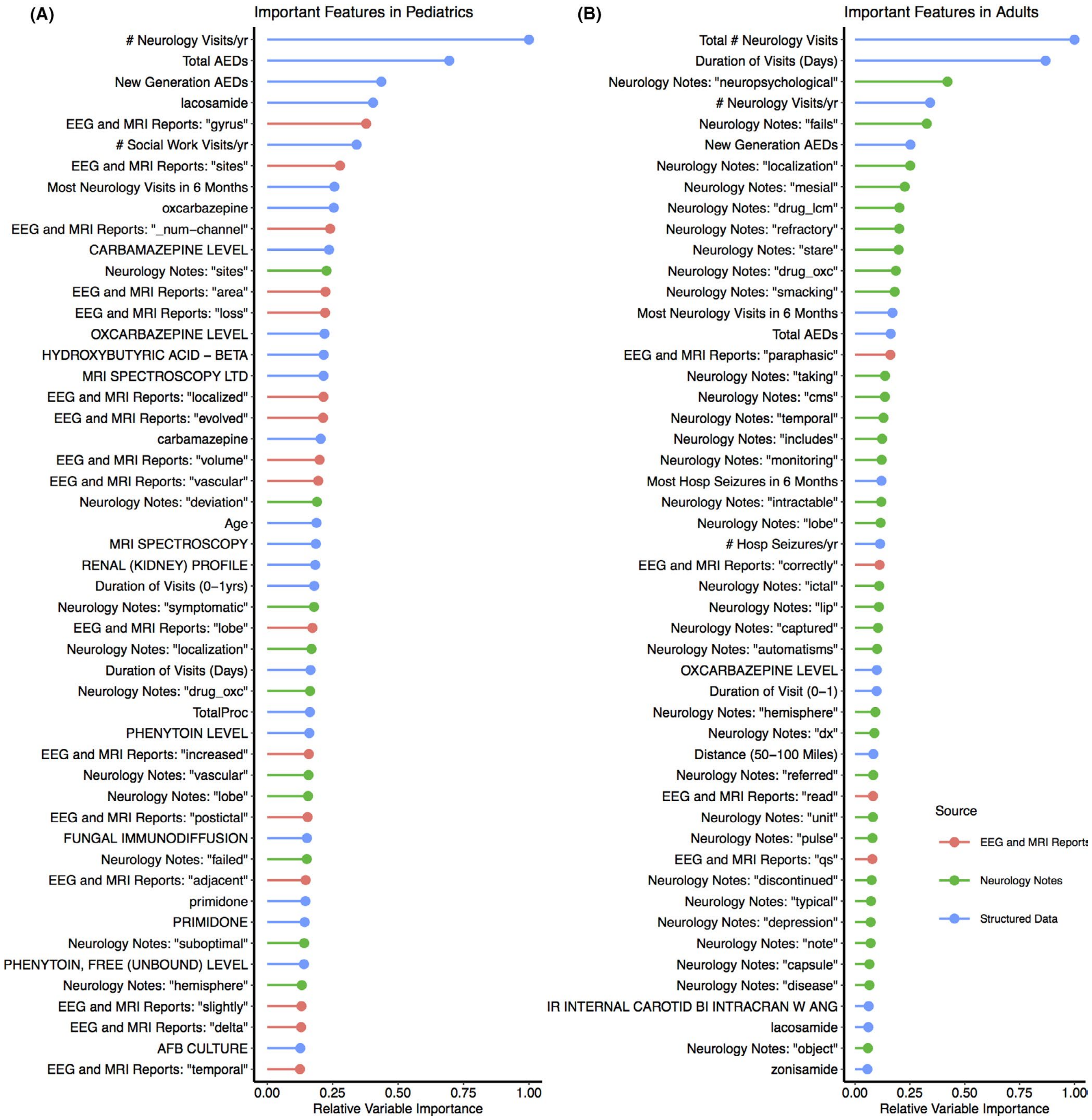

FIGURE 2 Top 50 features from the electronic health record. The pediatric dataset is shown in (A), and the adult dataset is shown in (B). Features in all capital letters correspond to orders for laboratories or procedures. Medications in lower case letters correspond to prescription orders. AED, anti-epileptic drug; drug_oxc, oxcarbazepine; drug_lcm, Lacosamide

their first seizure, while adult patients usually present after having epilepsy for many years. This may explain why the ML models were able to identify adult surgical patients at their second visit more accurately than in pediatrics. PR-AUC was low at both centers because, unlike AUROC, PR-AUC is dependent on population prevalence. It increases with increasing prevalence and decreases with decreasing prevalence. In this case, the prevalence of patients with epilepsy who undergo resective surgery is rare: 2.3% in the pediatric cohort and 0.7% in the adult cohort. This explains why the algorithm had a higher PR-AUC in the pediatric cohort, despite having a lower AUROC. PR-AUC was calculated using surgery as the outcome. Our previous study showed that the prevalence of surgical candidates in the "non-surgical" cohort at the pediatric center was 7%.[13] This means that some of the "false positives" were correct classifications. Therefore, the PR-AUC reported here is likely an underestimate.

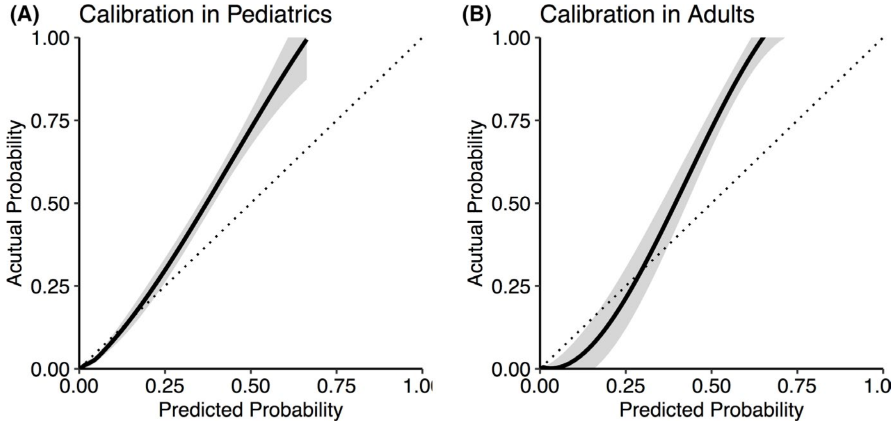

FIGURE 3 Calibration curves for the random forest. The pediatric dataset is on the left (A) and the adult dataset is on the right (B). Perfect calibration is denoted by the gray, dotted line, $y = x$. “predicted probability” is the model's estimated likelihood that the patient was a surgical candidate. For a given value of “predicted probability,” the “actual probability” is the observed proportion of patients who underwent surgical treatment. In (A), the average difference between the calibration curve and the identity line for all patients was 0.6%, the maximum difference was 33%, and the 90% quantile was 1.6%. For (B), the average difference between the calibration curve and the identity line was 0.5%, the maximum difference was 36%, and the 90% quantile was 1.3%

4.1 | Strengths of the study

The site-specific models presented here performed better than the existing neurology notes model.[13] We compared the models' performance against baseline models that included demographic information, follow-up duration, and the number of AEDs. Using these criteria alone were not enough to accurately predict which patients underwent surgery. The multimodal models had access to larger and more granular information about patients' medical history, which lead to more discriminative surgical candidacy scores. We demonstrated that this ML methodology for EHR-based feature extraction and selection could be generalized from pediatric to adult populations across two distinct institutions. It is not yet possible to apply one universal model at all epilepsy centers. Once EHR data from a sufficient sample of epilepsy centers are aggregated, it is possible that one generalizable model could be developed. Our methodology incentivizes centers to contribute toward such an effort. Site-specific algorithms can be used in the interim.

4.2 | Limitations of the study

While the datasets came from two distinct healthcare systems, there may have been a geographic selection bias since both are located in the Cincinnati metropolitan area. This methodology needs to be validated in additional regions. Temporal validation on a held-out set of new patient visits is also needed. In concordance with TRIPOD guidelines, this study used cross-validation instead of one held-out set of patients due to the relatively limited number of surgical patients. Although the algorithms identified surgical candidates before they entered the presurgical evaluation protocol, it is unclear whether the algorithms identified potential candidates before providers. Providers know which patients may be likely to require surgical treatment and, consciously or subconsciously, change the language of their notes accordingly. These subtle differences in the notes were used to quantify the likelihood of surgical candidacy. Finally, deep learning methods have been developed to classify patients using multimodal EHR data.[35,36] In certain situations, deep learning outperforms traditional ML algorithms used here, but they require thousands of cases for training.[37] We, like most other medical ML problems, had an order or magnitude fewer cases than would have been required to take advantage of these deep learning techniques. Our method provides a framework for identifying a limited number of features from EHR-wide data that can be used to classify patients with excellent performance.

5 | CONCLUSIONS

In conclusion, this methodology provided a means to produce site-specific ML models that could be used to identify surgical candidates years before their presurgical evaluations. Replication of strong results in two distinct patient populations indicates the potential to generalize this modeling procedure to additional epilepsy centers. After validating this approach prospectively, surgical candidacy scores could be sent to clinicians to provide decision support that facilitates earlier referrals for surgery.

ACKNOWLEDGEMENTS

Research reported in this manuscript was supported by the National Institutes of Health (F31 NS115447 and K25 HL125954) and the Agency for Healthcare Research and Quality (R21 HS024977). The authors would like to thank the numerous staff at Cincinnati Children's Hospital Medical Center's Division of Biomedical Informatics and the University of Cincinnati's Center for Health Informatics for providing the infrastructure needed to conduct this study; we also thank the patients with epilepsy who

were included in this study and hope these efforts help to improve future epilepsy care.

CONFLICT OF INTEREST

The authors have no competing interests to declare. Drs. Greiner, Glauser, and Pestian report a patent for the identification of surgery candidates using natural language processing (application num. 16/396 835), licensed to Cincinnati Children's Hospital Medical Center. All other authors report no disclosures.

AUTHOR CONTRIBUTION

B.D.W. contributed to the study design, data collection, data cleaning, machine learning analysis, interpretation of results, and wrote the first draft of the paper. H.M.G., T.A.G., J.P.P., D.S., D.M.F., F.T.M., and J.W.D. contributed to study design, interpretation of results, and critical review of the paper draft. A.J.K. contributed to data cleaning, machine learning analysis, and critical review of the paper draft. R.D.S. contributed to study design, statistical analysis, interpretation of results, and critical review of the paper draft.

ORCID

Benjamin D. Wissel https://orcid.org/0000-0002-8967-9296
Tracy A. Glauser https://orcid.org/0000-0003-1520-2732
Daniel Santel https://orcid.org/0000-0002-6495-8328

REFERENCES

1. Beghi E, Giussani G, Nichols E, et al. Global, regional, and national burden of epilepsy, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol*. 2019;*18*(4):357-375.
2. Kwan P, Schachter SC, Brodie MJ. Drug-resistant epilepsy. *N Engl J Med*. 2011;*365*(10):919-926.
3. Jobst BC, Cascino GD. Resective epilepsy surgery for drug-resistant focal epilepsy: a review. *JAMA*. 2015;*313*(3):285-293.
4. Engel J Jr, McDermott MP, Wiebe S, et al. Early surgical therapy for drug-resistant temporal lobe epilepsy: a randomized trial. *JAMA*. 2012;*307*(9):922-930.
5. Lamberink HJ, Otte WM, Blümcke I, et al. Seizure outcome and use of antiepileptic drugs after epilepsy surgery according to histopathological diagnosis: a retrospective multicentre cohort study. *Lancet Neurol*. 2020;*19*(9):748-757.
6. Englot DJ, Ouyang D, Garcia PA, Barbaro NM, Chang EF. Epilepsy surgery trends in the United States, 1990–2008. *Neurology*. 2012;*78*(16):1200.
7. Choi H, Sell RL, Lenert L, et al. Epilepsy surgery for pharmacoresistant temporal lobe epilepsy: a decision analysis. *JAMA*. 2008;*300*(21):2497-2505.
8. Sperling MR, Barshow S, Nei M, Asadi-Pooya AA. A reappraisal of mortality after epilepsy surgery. *Neurology*. 2016;*86*(21):1938-1944.
9. Téllez-Zenteno JF, Dhar R, Hernandez-Ronquillo L, Wiebe S. Long-term outcomes in epilepsy surgery: antiepileptic drugs, mortality, cognitive and psychosocial aspects. *Brain*. 2006;*130*(2):334-345.
10. Langfitt JT, Holloway RG, McDermott MP, et al. Health care costs decline after successful epilepsy surgery. *Neurology*. 2007;*68*(16):1290-1298.
11. Roberts JI, Hrazdil C, Wiebe S, et al. Neurologists' knowledge of and attitudes toward epilepsy surgery: a national survey. *Neurology*. 2015;*84*(2):159-166.
12. Cohen KB, Glass B, Greiner HM, et al. Methodological issues in predicting pediatric epilepsy surgery candidates through natural language processing and machine learning. *Biomed Inform Insights*. 2016;*8*:11-18.
13. Wissel BD, Greiner HM, Glauser TA, et al. Prospective validation of a machine learning model that uses provider notes to identify candidates for resective epilepsy surgery. *Epilepsia*. 2020;*61*(1):39-48.
14. Matykiewicz P, Cohen K, Holland KD, et al. Earlier identification of epilepsy surgery candidates using natural language processing. Proceedings of the 2013 Workshop on Biomedical Natural Language Processing; 2013.
15. Wissel BD, Greiner HM, Glauser TA, et al. Investigation of bias in an epilepsy machine learning algorithm trained on physician notes. *Epilepsia*. 2019;*60*(9):e93-e98.
16. Cross JH, Jayakar P, Nordli D, et al. Proposed criteria for referral and evaluation of children for epilepsy surgery: recommendations of the Subcommission for Pediatric Epilepsy Surgery. *Epilepsia*. 2006;*47*(6):952-959.
17. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digital Health*. 2020;*2*(9):e489-e492.
18. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg*. 2015;*102*(3):148-158.
19. Jetté N, Reid AY, Quan H, Hill MD, Wiebe S. How accurate is ICD coding for epilepsy? *Epilepsia*. 2010;*51*(1):62-69.
20. Benoit K, Watanabe K, Wang H, et al. Quanteda: an R package for the quantitative analysis of textual data. *J Open Source Softw*. 2018;*3*(30):774.
21. Grinspan ZM, Shapiro JS, Abramson EL, Hooker G, Kaushal R, Kern LM. Predicting frequent ED use by people with epilepsy with health information exchange data. *Neurology*. 2015;*85*(12):1031-1038.
22. Bautista RE, Glen ET, Wludyka PS, Shetty NK. Factors associated with utilization of healthcare resources among epilepsy patients. *Epilepsy Res*. 2008;*79*(2–3):120-129.
23. Jette N, Quan H, Tellez-Zenteno JF, et al. Development of an online tool to determine appropriateness for an epilepsy surgery evaluation. *Neurology*. 2012;*79*(11):1084-1093.
24. Berg AT, Vickrey BG, Langfitt JT, et al. The multicenter study of epilepsy surgery: recruitment and selection for surgery. *Epilepsia*. 2003;*44*(11):1425-1433.
25. Zip Code Characteristics. Mean and Median Household Income. 2010; https://www.psc.isr.umich.edu/dis/census/Features/tract2zip/. Accessed February 26, 2020.
26. Schiltz NK, Kaiboriboon K, Koroukian SM, Singer ME, Love TE. Long-term reduction of health care costs and utilization after epilepsy surgery. *Epilepsia*. 2016;*57*(2):316-324.
27. Nembrini S, König IR, Wright MN. The revival of the Gini importance? *Bioinformatics (Oxford, England)*. 2018;*34*(21):3711-3718.
28. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:150804409. 2015.
29. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol*. 2015;*68*(8):855-859.
30. Saito T, Rehmsmeier M. Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics*. 2017;*33*(1):145-147.
31. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med*. 2014;*33*(3):517-535.
32. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;*78*(1):1-3.

33. Team RC. R: a language and environment for statistical computing. 2013.
34. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. Reply. *N Engl J Med*. 2019;*380*(26):2589-2590.
35. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*. 2018;*1*(1):18.
36. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform*. 2017;*22*(5):1589-1604.
37. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2018;*25*(10):1419-1428.

SUPPORTING INFORMATION

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

How to cite this article: Wissel BD, Greiner HM, Glauser TA, et al. Early identification of epilepsy surgery candidates: A multicenter, machine learning study. *Acta Neurol Scand*. 2021;144:41–50. https://doi.org/10.1111/ane.13418