HHS Public Access

Author manuscript

Epilepsia. Author manuscript; available in PMC 2024 March 01.

Published in final edited form as:

Epilepsia. 2023 June ; 64(6): 1472–1481. doi:10.1111/epi.17589.

Identification of patients with epilepsy using automated electronic health records phenotyping

Marta Fernandes[1,2,3,4], **Aidan Cardall**[1,2,3], **Jin Jing**[1,2,3,4], **Wendong Ge**[1,2,3,4], **Lidia M. V. R. Moura**[1,2], **Claire Jacobs**[1,2], **Christopher McGraw**[1,2], **Sahar F. Zafar**[1,2], **M. Brandon Westover**[1,2,3,4]

[1]Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA

[2]Harvard Medical School, Boston, Massachusetts, USA

[3]Clinical Data Animation Center, Massachusetts General Hospital, Boston, Massachusetts, USA

[4]Henry and Allison McCance Center for Brain Health, Massachusetts General Hospital, Boston, Massachusetts, USA

Abstract

Objective: Unstructured data present in electronic health records (EHR) are a rich source of medical information; however, their abstraction is labor intensive. Automated EHR phenotyping (AEP) can reduce the need for manual chart review. We present an AEP model that is designed to automatically identify patients diagnosed with epilepsy.

Methods: The ground truth for model training and evaluation was captured from a combination of structured questionnaires filled out by physicians for a subset of patients and manual chart review using customized software. Modeling features included indicators of the presence of keywords and phrases in unstructured clinical notes, prescriptions for antiseizure medications (ASMs), International Classification of Diseases (ICD) codes for seizures and epilepsy, number of ASMs and epilepsy-related ICD codes, age, and sex. Data were randomly divided into training (70%) and hold-out testing (30%) sets, with distinct patients in each set. We trained regularized logistic regression and an extreme gradient boosting models. Model performance was measured using area under the receiver operating curve (AUROC) and area under the precision–recall curve (AUPRC), with 95% confidence intervals (CI) estimated via bootstrapping.

Correspondence: Marta Fernandes, 55 Fruit Street, Boston, MA 02114, USA. mbentofernandes@mgh.harvard.edu.
Marta Fernandes and Aidan Cardall are co-first authors.
Sahar F. Zafar and M. Brandon Westover are co-senior authors.

AUTHOR CONTRIBUTIONS

Study conception and design: Marta Fernandes, Aidan Cardall, Lidia M. V. R. Moura, Claire Jacobs, Christopher McGraw, Sahar F. Zafar, M. Brandon Westover. Data curation: Marta Fernandes, Aidan Cardall, Sahar F. Zafar, M. Brandon Westover. Data annotation: Marta Fernandes, Aidan Cardall. Data acquisition: Marta Fernandes, Jin Jing, Wendong Ge. Data processing: Marta Fernandes, Aidan Cardall. Analysis and interpretation of data: Marta Fernandes, Aidan Cardall, M. Brandon Westover. Drafting/revision of the manuscript for content, including medical writing: Marta Fernandes, Aidan Cardall, Sahar F. Zafar, M. Brandon Westover.

CONFLICT OF INTEREST STATEMENT

M.B.W. is a cofounder of Beacon Biosignals, which played no role in this work. S.F.Z. is a clinical neurophysiologist for Corticare, which played no role in this work. All other authors report no potential conflicts of interest. 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.

SUPPORTING INFORMATION

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

Results: Our study cohort included 3903 adults drawn from outpatient departments of nine hospitals between February 2015 and June 2022 (mean age = 47 ± 18 years, 57% women, 82% White, 84% non-Hispanic, 70% with epilepsy). The final models included 285 features, including 246 keywords and phrases captured from 8415 encounters. Both models achieved AUROC and AUPRC of 1 (95% CI = .99–1.00) in the hold-out testing set.

Significance: A machine learning-based AEP approach accurately identifies patients with epilepsy from notes, ICD codes, and ASMs. This model can enable large-scale epilepsy research using EHR databases.

Keywords

electronic medical records (EMR); neurology; text mining; unstructured text

1 | INTRODUCTION

Electronic health records (EHR) contain information useful for clinical and research applications.[1] However, >80% of EHR data are in unstructured clinical notes.[2,3] Notes data are difficult to analyze at scale because they typically require human chart review.[4–6]

Automated EHR phenotyping (AEP), in which features from notes are combined with structured information, has emerged as a solution[7] to facilitate research in large EHR datasets. AEP has value in identifying diagnoses,[8] especially for prevalent diseases. Epilepsy affects 50–70 million people worldwide[9,10] and is associated with high costs,[11] premature death, and lost work.[9,12,13] Prior studies have used diagnosis codes from the International Classification of Diseases (ICD-9, ICD-10) to identify patients with epilepsy.[14–20] However, ICD codes are intended primarily for billing purposes rather than communicating medical diagnoses, and inferring diagnoses from ICD codes can be inaccurate.[8] Although current applications of AEP in epilepsy have so far been limited by algorithm accuracy and generalizability, larger study sizes in genetic studies and precision medicine trials are making AEP essential for epilepsy phenotyping.[21]

We present a machine learning-based AEP approach to identifying patients with epilepsy from unstructured clinical notes, ICD codes, antiseizure medications (ASMs), and demographics from EHR.

2 | MATERIALS AND METHODS

2.1 | Study cohort

This study is reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement.[22] EHR data were extracted under a protocol approved by the institutional review board with a waiver of informed consent. Data were from adult patients (age ≥ 18 years) seen in epilepsy clinics and other departments.

The ground truth for the diagnosis of epilepsy was extracted using one of four methods: first, when available, we extracted the ground truth directly from structured questionnaires (available for a subset of patients[23] as part of a quality improvement process; $n = 2277$

patients, $N = 6916$ visits); second, by computationally searching for a small set of phrases that clearly stated a patient did not have a diagnosis of epilepsy (specifically, "no epilepsy", "not epilepsy", or "does not have epilepsy") followed by manual verification ($n = 42$, $N = 45$); third, by manual review of charts using the EHR interface (Epic; $n = 13$, $N = 69$); and fourth, by manual review using customized software (Prodigy24; we made our code for annotations with Prodigy publicly available, including the code for sentence tokenization, pattern matching from a list of keywords presented in Table S1, and prompts to run the application programming interface [https://github.com/mpriscila88/epilepsy_classification]; for patients seen in the epilepsy clinics: $n = 911$, $N = 1000$; other departments: n, $N = 1500$). Notes selected for annotation[24] were randomly selected from epilepsy clinics ($N = 1000$) between January 12, 2016 and June 14, 2022, and other from departments ($N = 1500$) from February 1, 2015 to June 14, 2022. Cases lacking a ground truth diagnosis (diagnosis of epilepsy uncertain) were excluded. After excluding uncertain cases ($n = 736$, $N = 1103$) and nonadults ($n = 2$, $N = 12$), the cohort included 3903 patients with 8415 notes, where 69% ($n = 2708$, $N = 7090$) had a diagnosis of epilepsy. There were 2741 distinct patients seen in outpatient epilepsy clinics; the remaining 1162 were seen in other departments.

2.2 | Features for modeling

The initial set of text features included 300 keywords and phrases (Table S1). Of these, 246 were present in the notes of our dataset (Table S2). The final set of features included age, sex, 32 groups of ASMs, four groups of ICD codes, the number of active ASMs, number of ICD codes assigned at the visit, and 246 text features, for a total of 286 features. ICD groupings and ASMs were defined a priori by two epileptologists (S.F.Z., M.B.W.). All features were binary except age, number of ICD codes, and number of ASMs. ICD codes and ASM groupings, and text features and their preprocessing, were as follows:

- ICD codes: ICD codes were grouped into categories: "epilepsy and recurrent seizures"—ICD-10G40.* and ICD-9345.* codes, plus genetic epilepsy codes Q04.3, R56.9, 742.4, and 780.39; "convulsions/seizures"— ICD-10 R56.* and ICD-9780.39; and syncope— ICD-10 R55 and ICD-9780.2 codes. A day prior to and after the visit was considered for assignment of an ICD code to account for prior or delayed data entry.
- ASMs: ASMs included acetazolamide, brivaracetam, cannabidiol, carbamazepine, cenobamate, clobazam, clonazepam, clorazepate, diazepam, eslicarbazepine, ethosuximide, ezogabine, felbamate, gabapentin, ketamine, lacosamide, lamotrigine, levetiracetam, lorazepam, methsuximide, midazolam, oxcarbazepine, perampanel, phenobarbital, phenytoin, pregabalin, primidone, rufinamide, tiagabine, topiramate, valproic acid, and zonisamide.
- Text features: Features were extracted from notes after preprocessing. Special characters and duplicated blank spaces were removed, followed by lowercasing. Text was stemmed using SnowballStemmer, an updated version of Porter stemmer,[25] where each word is reduced to its base word or stem (e.g., "cares", "cared" ➔ "care"). Features from notes were indicators of the presence of keywords and phrases, or "bags of words" (unordered sets of keywords), defined by the study team (Table S1). Notes were tokenized at the sentence

level. Sentences were extracted for further featurization if they included one or more of the keywords or phrases. The remaining text was not used. A binary feature column indicating presence or absence of each phrase or "bag" of keywords was created. For example, for "I do not recommend antiepileptic medication", the bag of (stemmed) keywords consisted of {"not", "recommend", "antiepilept", "medic"}. In case a sentence containing these words was present, the corresponding binary feature column was assigned "1" for this note. Text features from the initial set were automatically excluded if not present in any notes. Groups of features that were semantically equivalent, (e.g., "seizure free" and "sz free"), were merged into a single binary column. A list of merged bag-of-words features is presented in Table S3.

2.3 | Modeling design

Our model consists of two stages. The first applies a simple rule based on ICD codes and ASMs to classify "easy cases." For cases not classified as "easy" (i.e., "hard cases"), the model applies a more complex machine learning algorithm (see below).

2.3.1 | Stage 1: Classification of "easy" cases—The model first categorizes a patient into groups based on whether they have any epilepsy-related ICD code (ICD+) versus none (ICD−), and whether patients had been prescribed any ASMs (ASM+) versus none (ASM−). We defined four groups. "Easy cases" are defined as ASM+ ICD+, ASM− ICD−, in which ASM and ICD status are concordant, and "hard cases" as ASM− ICD+, ASM+ ICD−, in which ASM and ICD status are discordant. Easy cases are categorized by the model as follows: patients in the ASM− ICD− ("easy negative") group are classified a priori as not having epilepsy; those in the ASM+ ICD+ ("easy positive") group are classified as having epilepsy. Patients not in the group of easy cases are classified by Stage 2 of the model.

2.3.2 | Stage 2: Classification of "hard" cases—For patient visits in the mixed groups (ASM+ ICD−, ASM− ICD+), we developed a machine learning model (Stage 2 of the classification model). To develop the model, cases were divided randomly to create a dataset with distinct patients in train (70%) and hold-out test (30%) sets. Age, number of ICD codes, and number of ASMs were normalized using minimum–maximum normalization.[26] Hyperparameter tuning was performed to train a regularized logistic regression (LR) model[27] and an extreme gradient boosting (XGBoost)[28] model in 100 iterations of fivefold cross-validation using the training data. Hyperparameters selected for tuning are presented in Appendix S1.1. Stage 2 models were created including five sets of features (total number): text features (246); ICD codes, including number of ICD codes assigned (4); ASMs, including number of ASMs prescribed (33); ICD codes, ASMs, age, sex (39); and text features, ICD codes, ASMs, age, sex (285). Performance of the models in the test set was compared. For LR, the relative importance of features was assessed by magnitude of the regression coefficients; for XGBoost, we used SHAP (Shapley Additive Explanations) values.[29]

2.4 | Performance evaluation

Model performance was evaluated using area under the precision–recall and receiver operating characteristic curves (AUPRC, AUROC).[30] AUPRC quantifies the trade-off between precision (i.e., positive predictive value) and recall (i.e., sensitivity). AUROC quantifies the trade-off between sensitivity and false positive rate[31]. We selected a threshold to convert model output probabilities to binary decisions (for Stage 2 of the model; Stage 1 already produces binary decisions for easy cases). The threshold was calculated with training data based on maximization of the F1 score (harmonic mean of precision and recall; F1 = [precision × recall]/[precision + recall]), which is suited for imbalanced data.[30] These binary decisions are compared to the ground truth labels, and categorized as true and false positives, and true and false negative classifications, from which we calculated model specificity and F1 score.[32] For each metric, we present the macro average for both positive and negative diagnoses of epilepsy. That is, we calculate the metric with epilepsy as the target class, and with nonepilepsy as the target class, and compute the average. We performed 1000 bootstrapping iterations to calculate 95% confidence intervals (CIs).

3 | RESULTS

3.1 | Patient characteristics

Our final cohort comprised 3903 patients with 8415 visits (see CONSORT[33] diagram in Figure 1). Most patients were women (57.5%), White (82.3%), and non- Hispanic (83.9%), with an average age of 46 years at baseline (Table 1). A total of 2733 (70%) patients had a diagnosis of epilepsy.

Patient visits with a documented diagnosis of epilepsy were prescribed a median of 2 (interquartile range [IQR] = 1–3) outpatient ASMs, whereas patients with a negative diagnosis were prescribed 0 (IQR = 0–1). Lamotrigine (37.2%) was the most prescribed ASM, followed by levetiracetam (36.5%), lorazepam (25.3%), and lacosamide (14.1%; Table S4). The most commonly assigned epilepsy-related ICD codes were “epilepsy and recurrent seizures” (ICD-10G40.* and ICD-9345.*; 63.7%). Text features that appeared most often in patients with epilepsy included “history of seizures”, “with epilepsy”, “focal”, “lamotrigine”, and “keppra”.

3.2 | Model performance

3.2.1 | Performance on “easy” cases—The test set included 1983 “easy” visits (942 patients). Among these, 160 (155 patients) had neither epilepsy ICD codes nor ASMs (ASM− ICD− group), and 1823 (787 patients) had both ICD codes and ASMs (ASM+ ICD+). Stage 1 of the model classifies all ASM− ICD− visits as not having epilepsy, and all ASM+ ICD+ patient visits as having epilepsy; thus, the model will be incorrect for any ASM− ICD− cases that have a diagnosis of epilepsy (false negative), and for ASM+ ICD+ cases that do not have a diagnosis of epilepsy (false positive).

Within the ASM− ICD− group, no visits had a diagnosis of epilepsy, yielding a false negative rate of 0% (0/160).

Within the ASM+ ICD+ group, 30 visits were not associated with an epilepsy diagnosis, yielding a false positive rate of 1.6% (30/1823). The overall error rate (false positives + false negatives) on easy cases was 1.5% (30/1983). Upon chart review, we identified four types of reasons for the presence of epilepsy ICD codes and ASMs (number of visits): (1) patient had been diagnosed with epilepsy and prescribed ASM by a prior neurologist but was being weaned off by a new neurologist (19); (2) patient had experienced a single provoked seizure that was not felt to be epilepsy, but was taking an ASM for a psychiatric indication or migraines (6); (3) patient was prescribed ASM for suspected epilepsy but had not been given a diagnosis (3); and (4) patient had a previous diagnosis of epilepsy but was now considered in remission and was tapering off ASMs (2).

3.2.2 | Performance on "hard" cases—We next evaluated performance of Stage 2, the machine learning algorithm that classifies "hard" cases (ASM+ ICD−, ASM− ICD+). Overall, among hard test set cases there were 547 visits (404 patients), among which 72.0% (394/547) had a diagnosis of epilepsy, and 28.0% (153/547) did not. The ASM+ ICD− group had 491 patients, 79.8% (392/491) with an epilepsy diagnosis; the ASM− ICD+ group had 56 patients, 4.6% (2/56) with an epilepsy diagnosis. We report performance of several versions of the Stage 2 model to analyze the importance of the classification model type (simple linear model [LR] vs. complex model [XGBoost]), and type of information provided to the model (ICD codes vs. ASMs vs. text-based features vs. combinations).

Using all features (ASM, ICD, text), performance of LR and XGBoost models was similar (Table 2). AUROC and AUPRC curves for the LR model are shown in Figure S2; confusion matrices are in Figure S3. Hyperparameters selected in fivefold cross-validation are shown in Table S5. Model performance within the four groups is summarized in Table S6. Because the two models performed similarly, from here onward we discuss only the LR results, as this simple model is more interpretable.

To investigate importance of various types of information, we trained the LR model with the following: ICD codes alone, ASMs alone, text-based features alone, all non-text-based features, and all features combined (Figure 2). With ICD codes alone the model performed poorly, with a recall of .51 (95% CI = .50–.51) and low F1 score of .44 (95% CI = .42–.47). With ASMs, the model performed better, with recall of .90 (95% CI = .87–.93) and F1 score of .91 (95% CI = .88–.94). After adding ICD codes, age, and sex to ASMs, performance was similar, with a slight decrease of 1% in AUPRC, whereas AUROC remained unchanged. Performance with text features alone surpassed performance using all features except text. Combining all features, AUPRC and AUROC had a 1% increase compared to text only, whereas recall and F1 score increased 3% and 2%. Models with all feature types performed best, with macro average AUROC and AUPRC of 1 (95% CI = .99–1), as depicted in Table 2. From here on "LR model" refers to the LR model trained with all features.

Within the 547 hard cases in the test data, the LR model misclassified eight visits; thus, the overall error rate for hard cases was 8/547 = 1.46%. All eight misclassifications occurred within the ASM+ ICD− group; six patients with epilepsy were incorrectly classified as not having epilepsy (false negatives), yielding a false negative rate of 6/392 = 1.53%; two patients without epilepsy were incorrectly classified as having epilepsy (false positives),

yielding a false positive rate of 2/99 = 2.02%. Within the ASM− ICD+ group, all 54 patients who did not have epilepsy and two who had the diagnosis were correctly classified.

These mixed cases within the ASM+ ICD− group are more ambiguous than the "easy cases," lacking either an ASM prescription or an ICD code. On manual chart review, reasons for false positives in this group were a patient with prior seizures but no diagnosis of epilepsy taking an ASM for a nonseizure indication (migraines), and a patient with a prior seizure disorder now off ASMs and no longer considered to have epilepsy. Reasons for false negatives included histories of psychiatric illness,[2] attention deficit hyperactivity disorder,[1] obsessive compulsive disorder,[1] mood disorders,[1] panic and anxiety,[1] depression,[1] and complaints related to pain[1] in patients who also had a diagnosis of epilepsy. In these cases, emphasis in physicians' notes was on the nonepilepsy diagnosis, and identifying that the patient had a diagnosis of epilepsy was difficult even for chart reviewers.

3.2.3 | Overall performance in the test set—The overall model, evaluated on the full test set (including both easy and hard cases), achieved a macro average AUROC and AUPRC of 1.00 (95% CI = .99–1; Table S7). AUROC and AUPRC curves are shown in Figure S1.

Overall performance within the general hospital population: Prevalence of epilepsy within the overall hospital population is smaller than in our test set. To estimate the expected performance in the general hospital population, we selected an additional random sample of 1000 adults from the EHR, to estimate the prevalence of patients within each of the four groups. These percentages and numbers are as follows: ASM+ ICD+, .1% ($n = 1$); ASM+ ICD−, .1% ($n = 1$); ASM− ICD+, 4.1% ($n = 41$); ASM− ICD−, 95.7% ($n = 957$). Putting these prevalence values (p^{++}, p^{+-}, p^{-+}, p^{--}) together with the corresponding model error rates (Pe^{++}, Pe^{+-}, Pe^{-+}, Pe^{--}), performance within the overall hospital population is estimated to be:

$$P[E] = Pe^{++}p^{++} + Pe^{\pm}p^{\pm} + Pe^{\mp}p^{\mp} + Pe^{--}p^{--} = .012 \cdot .01 + .012 \cdot .01 + 0 \cdot .04 + 0 \cdot .96 = .0003$$

The overall error rate in the general (hospital) population is thus .03%, which is quite small.

3.3 | Features importance

The relative importance of the top 20 features in the final LR model is presented in Figure 3 (for the XGBoost model, see Figure S4). Prescription of lamotrigine was the most important feature, followed by levetiracetam. Prescription of other ASMs (lacosamide, valproic acid, carbamazepine, clobazam) were among the top 20. The number of ASMs prescribed was also important, as well as the group of ICD codes for "epilepsy and recurrent seizures." The bag of words feature {"partial", "seizur"} referring to partial seizure(s) was the third most important feature. Reference to "tonic clonic seizures" or "sudden unexpected death" were also important. Features contributing to classification of a negative diagnosis of epilepsy included reference to not taking ASMs ({"no", "antiseizur", "medic"}); no epileptiform activity ({"not", "epileptiform", "activ"}); the keyword "psychogenic", which was often associated with mentions of psychogenic nonepileptic seizures (PNES), and psychiatric and

psychological factors such as stress, depression, or anxiety; acute symptomatic seizures ({“acut”, “symptomat”, “seizur”}; as opposed to epilepsy[34]); and the keyword “vasovagal”, which relates to fainting.[35]

4 | DISCUSSION

4.1 | Principal findings

A machine learning-based AEP approach achieved excellent performance for identifying patients with a diagnosis of epilepsy from EHR data by integrating information from clinical notes, ICD codes, and prescriptions for ASMs. The linear model (LR) achieved similar results to the black-box model (XGBoost), with the advantage of being simpler and more easily interpretable. The model substantially outperformed models that relied only on ICD codes, only ASMs, or a combination of both. The model developed in this work allows identification of patients with epilepsy from the EHR at scale, paving the way for large-scale EHR-based studies of epilepsy.[12,13]

We observed that the AEP model using text features alone exhibited very high performance, nearly as high as the model using all features. For the group of patient visits with indication of an ASM or an epilepsy-related ICD code, the LR model using all features achieved an AUROC of 1 (95% CI = .99–1.00) and an AUPRC of 1 (95% CI = .99–1.00) and .99 (95% CI = .98–1.00) for positive and negative diagnosis of epilepsy, respectively. The most important features contributing to the classification consisted of prescription of certain ASMs, with lamotrigine ranked most important, ICD codes for “epilepsy and recurrent seizures,” and text features. Some text features contributing to the positive diagnosis of epilepsy were related to types of seizures, whereas the most prominent features for the negative diagnosis were related to not taking or needing to take ASM, having no epileptiform activity on electroencephalographic (EEG) testing, the presence of psychogenic factors, or text indicating acute symptomatic seizures and vasovagal episodes.

4.2 | Comparison with prior work

Several prior studies have investigated approaches to identifying epilepsy using EHR data mining and varying combinations of clinical notes, ICD codes, EEG, and ASM utilization.[14–20,36] The reported precision and recall of prior work has ranged from .33 to 1.00, and .22 to 1.00, respectively.[14–20,36] Limitations of prior literature include small sample sizes.[15,18–20,36] Other limitations include overestimation of epilepsy cases, and higher number of false positives, likely because of not including data from clinical notes and therefore insufficiently capturing nonseizure indications for ASMs.[14] Prior work that has used structured epilepsy questionnaires limited their analysis only to patients with completed questionnaires and those seen by a neurologist.[15] In contrast, we included varying provider types (epileptologists, other neurologists, and nonneurologists) to increase generalizability. Similarly, in another study using ICD codes alone, the model was developed only for patients seen in neurology clinics, where the prevalence of epilepsy is much higher than in a general patient population, likely inflating the estimated precision of the algorithms.[19] In other work using ASM prescriptions, pharmacy data were obtained only for patients who had prescriptions filled within pharmacies that contracted with a medical

plan, and prescriptions filled outside the system were unknown, which could have led to underestimated use of ASMs by patients with seizure disorders.[18] These factors might have affected the consistency of the data used for modeling and limited predictive power. In this study, the reliability and validity of using medical record reviews as a gold standard for establishing an epilepsy diagnosis depended in part on the diagnostic abilities of the health care providers and the consistency and quality of medical record keeping. In a systematic review[17] of 30 studies published between 1989 and 2018, the authors concluded that the precision was higher in algorithms combining ICD codes with ASMs. Our work shows that incorporating features derived from unstructured notes can further enhance the precision of models to identify epilepsy through EHR mining.

Other studies applied text mining techniques for epilepsy phenotyping from radiology reports,[21] clinician notes,[21] EEG reports,[21,37] and discharge summaries,[37–39] including identification of epilepsy syndromes,[21,37] EEG abnormalities,[21] and PNES.[21,40] Studies also applied text mining for classification of types of epilepsy from progress notes[41] and clinic letters[36] using ontology-based language;[42,43] epilepsy cohort generation[44] using standardized vocabularies; and extraction of abnormal findings on imaging, medications, and seizure frequency from clinic letters.[36] Our study is the first to combine text mining with ICD codes and medications, resulting in a model with higher accuracy in identification of epilepsy compared with prior work. Potential applications of our model include quality improvement and comparative effectiveness studies, and identification of patients for clinical trial eligibility.

4.3 | Limitations

The present analysis was limited to patients in a single hospital system, in one geographic region (Boston, Massachusetts). Thus, the cohort may not be representative of other US and non-US populations. Prescription of ASMs may also vary according to the patients' type of insurance, which was not considered in our modeling approach; thus, including this information in the model can be pursued in future work. Although findings from EEG and magnetic resonance imaging examinations might be present in notes, features from reports of these could be explicitly included for modeling in future work. Cases where the notes did not contain enough information to define a diagnosis, either due to lack of clinical information related to the diagnosis or brevity of some notes, were not considered in the study; however, in cases that we reviewed, this limitation appeared to be largely related to clinical diagnostic uncertainty on the part of clinicians regarding the ground truth rather than a limitation of our model. Although our model classifies the diagnosis, it does not classify the type of epilepsy, or provide information about the severity of epilepsy or seizure frequency. These other aspects of epilepsy are important determinants of life quality for patients with epilepsy; thus, developing models able to extract them is an important future direction.

5 | CONCLUSIONS

An interpretable machine learning-based natural language processing approach accurately identifies patients with epilepsy using a combination of clinical expertise in defining

keywords and phrases from unstructured clinical notes with ICD codes, antiseizure medications, and demographic patient data. This model will enable large-scale epilepsy research using EHR.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

During this research, M.B.W. was supported by the American Academy of Sleep Medicine through an AASM Foundation Strategic Research Award, and grants from the NIH (R01NS102190, R01NS102574, R01NS107291, RF1AG064312, RF1NS120947, R01AG073410, R01HL161253) and National Science Foundation (2014431). S.F.Z. was supported by the NIH (K23NS114201). L.M.V.R.M. was supported by the Centers for Diseases Control and Prevention (U48DP006377), the NIH (NIH-NIA 5K08AG053380–02, NIH-NIA 5R01AG062282–02, NIH-NIA 2P01AG032952–11, NIH-NIA 3R01AG062282–03S1, NIH-NIA 1R01AG073410–01), and the Epilepsy Foundation of America. C.J. received support from Harvard Catalyst | The Harvard Clinical and Translational Science Center (National Center for Advancing Translational Sciences, NIH award UL1 TR002541) and financial contributions from Harvard University and its affiliated academic health care centers. We also acknowledge the Epilepsy Learning Healthcare System. The funding sources had no role in study design, data collection, analysis, interpretation, or writing of the report. All authors had full access to all data, and the corresponding author had final responsibility for the decision to submit for publication.

Funding information

Centers for Disease Control and Prevention, Grant/Award Number: U48DP006377; NIH Clinical Center, Grant/Award Number: 1R01AG073410–01, 2P01AG032952–11, 3R01AG062282–03S1, 5K08AG053380–02, 5R01AG062282–02, K23NS114201, R01AG073410, R01HL161253, R01NS102190, R01NS102574, R01NS107291, RF1AG064312, RF1NS120947 and UL1 TR002541; National Science Foundation, Grant/Award Number: 2014431

REFERENCES

1. Baldassano SN, Hill CE, Shankar A, Bernabei J, Khankhanian P, Litt B. Big data in status epilepticus. Epilepsy Behav EB. 2019;101(Pt B):106457.
2. Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell J. Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. J Med Ethics. 2020;46(6):367–77. [PubMed: 32457202]
3. Kong HJ. Managing unstructured big data in healthcare system. Healthc Inform Res. 2019;25(1):1–2. [PubMed: 30788175]
4. Ford E, Nicholson A, Koeling R, Tate AR, Carroll J, Axelrod L, et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? BMC Med Res Methodol. 2013;13(1):105. [PubMed: 23964710]
5. Price SJ, Stapley SA, Shephard E, Barraclough K, Hamilton WT. Is omission of free text records a possible source of data loss and bias in clinical practice research datalink studies? A case-control study. BMJ Open. 2016;6(5):e011664.
6. Capurro D, Yetisgen M, van Eaton E, Black R, Tarczy-Hornoch P. Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: a multisite assessment. EGEMS Wash DC. 2014;2(1):1079. [PubMed: 25848594]
7. Wang Y, Wang L, Rastegar- Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform. 2018;77:34–49. [PubMed: 29162496]
8. HealthITAnalytics. Identifying disease with natural language processing technology [Internet]. HealthITAnalytics. 2021 [cited 2022 Sep 23]. Available from: https://healthitanalytics.com/news/identifying-disease-with-natural-language-processing-technology

9. Ngugi AK, Bottomley C, Kleinschmidt I, Sander JW, Newton CR. Estimation of the burden of active and life-time epilepsy: a meta-analytic approach. Epilepsia. 2010;51(5):883–90. [PubMed: 20067507]
10. Scheffer IE, Berkovic S, Capovilla G, Connolly MB, French J, Guilhoto L, et al. ILAE classification of the epilepsies: position paper of the ILAE Commission for Classification and Terminology. Epilepsia. 2017;58(4):512–21. [PubMed: 28276062]
11. Begley CE, Beghi E. The economic cost of epilepsy: a review of the literature. Epilepsia. 2002;43(Suppl 4):3–9.
12. Engel J ILAE classification of epilepsy syndromes. Epilepsy Res. 2006;70(Suppl 1):S5–10. [PubMed: 16822650]
13. Falco-Walter JJ, Scheffer IE, Fisher RS. The new definition and classification of seizures and epilepsy. Epilepsy Res. 2018;139:73–9. [PubMed: 29197668]
14. Bellini I, Policardo L, Zaccara G, Palumbo P, Rosati E, Torre E, et al. Identification of prevalent patients with epilepsy using administrative data: the Tuscany experience. Neurol Sci Off J Ital Neurol Soc Ital Soc Clin Neurophysiol. 2017;38(4): 571–7.
15. Jason RS, Felipe JSJ, Brandy EF, Jeffrey RB, Susan TH, Neishay A, et al. Accuracy of ICD-10-CM claims-based definitions for epilepsy and seizure type. Epilepsy Res. 2020;166:106414. [PubMed: 32683225]
16. Franchi C, Giussani G, Messina P, Montesano M, Romi S, Nobili A, et al. Validation of healthcare administrative data for the diagnosis of epilepsy. J Epidemiol Community Health. 2013;67(12):1019–24. [PubMed: 24022813]
17. Mbizvo GK, Bennett KH, Schnier C, Simpson CR, Duncan SE, Chin RFM. The accuracy of using administrative healthcare data to identify epilepsy cases: a systematic review of validation studies. Epilepsia. 2020;61(7):1319–35. [PubMed: 32474909]
18. Holden EW, Grossman E, Nguyen HT, Gunter MJ, Grebosky B, Von Worley A, et al. Developing a computer algorithm to identify epilepsy cases in managed care organizations. Dis Manag DM. 2005;8(1):1–14. [PubMed: 15722699]
19. Reid AY, St Germaine-Smith C, Liu M, Sadiq S, Quan H, Wiebe S, et al. Development and validation of a case definition for epilepsy for use with administrative health data. Epilepsy Res. 2012;102(3):173–9. [PubMed: 22727659]
20. Moura LMVR, Price M, Cole AJ, Hoch DB, Hsu J. Accuracy of claims-based algorithms for epilepsy research: revealing the unseen performance of claims-based studies. Epilepsia. 2017;58(4):683–91. [PubMed: 28199007]
21. Khankhanian P, Kosaraju N, Pathmanathan J, Ellis C, Helbig I, Litt B, et al. On the feasibility of natural language processing for standardized data extraction from electronic medical records of epilepsy patients (P1.283). Neurology. 2018;90(15 Supplement).
22. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. Int J Surg Lond Engl. 2014;12(12):1500–24.
23. Fernandes M, Donahue MA, Hoch D, Cash S, Zafar S, Jacobs C, et al. A replicable, open-source, data integration method to support national practice-based research & quality improvement systems. Epilepsy Res. 2022;18(186):107013.
24. Montani I, Honnibal M. Prodigy: a new tool for radically efficient machine teaching explosion [Internet]. Explosion. 2017. [cited 2022 Sep 20]. Available from: https://explosion.ai/blog/prodigy-annotation-tool-active-learning
25. Bird S, Klein E, Loper E. Natural Language Processing with Python [Internet]. 2019. [cited 2022 Sep 21]. Available from: https://www.nltk.org/book/
26. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Amsterdam: Elsevier; 2011. p. 740.
27. Cramer JS. The origins of logistic regression [Internet]. Rochester, NY: SSRN Electronic Journal; 2002. [cited 2022 Sep 21]. Available from: https://papers.ssrn.com/abstract=360300
28. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2016. p. 785–94. 10.1145/2939672.2939785

29. Lundberg S, Lee SI. A unified approach to interpreting model predictions [Internet]. arXiv. 2017 [cited 2022 Sep 14]. Available from: http://arxiv.org/abs/1705.07874
30. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS One. 2015;10(3):e0118432. [PubMed: 25738806]
31. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010 Jan;21(1):128–38. [PubMed: 20010215]
32. Azari A, Janeja VP, Levin S. Imbalanced learning to predict long stay Emergency Department patients. 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2015. p. 807–14.
33. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332. [PubMed: 20332509]
34. Sarmast ST, Abdullahi AM, Jahan N. Current classification of seizures and epilepsies: scope, limitations and recommendations for future action. Cureus. 2020;12(9):e10549. [PubMed: 33101797]
35. Jeanmonod R, Sahni D, Silberman M. Vasovagal Episode. StatPearls [Internet]. Treasure Island, FL: StatPearls Publishing; 2022 [cited 2022 Sep 20]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK470277/
36. Fonferko-Shadrach B, Lacey AS, Roberts A, Akbari A, Thompson S, Ford DV, et al. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system. BMJ Open. 2019;9(4): e023232.
37. Sullivan R, Yao R, Jarrar R, Buchhalter J, Gonzalez G. Text classification towards detecting misdiagnosis of an epilepsy syndrome in a pediatric population. AMIA Annu Symp Proc. 2014;2014:1082–7. [PubMed: 25954418]
38. Cui L, Sahoo SS, Lhatoo SD, Garg G, Rai P, Bozorgi A, et al. Complex epilepsy phenotype extraction from narrative clinical discharge summaries. J Biomed Inform. 2014;51: 272–9. [PubMed: 24973735]
39. Cui L, Bozorgi A, Lhatoo SD, Zhang GQ, Sahoo SS. EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification. AMIA Annu Symp Proc AMIA Symp. 2012;2012:1191–200. [PubMed: 23304396]
40. Hamid H, Fodeh SJ, Lizama AG, Czlapinski R, Pugh MJ, LaFrance WC, et al. Validating a natural language processing tool to exclude psychogenic nonepileptic seizures in electronic medical record-based epilepsy research. Epilepsy Behav EB. 2013;29(3):578–80.
41. Connolly B, Matykiewicz P, Bretonnel Cohen K, Standridge SM, Glauser TA, Dlugos DJ, et al. Assessing the similarity of surface linguistic features related to epilepsy across pediatric hospitals. J Am Med Inform Assoc. 2014;21(5):866–70. [PubMed: 24692393]
42. Kassahun Y, Perrone R, De Momi E, Berghöfer E, Tassi L, Canevini MP, et al. Automatic classification of epilepsy types using ontology-based and genetics-based machine learning. Artif Intell Med. 2014;61(2):79–88. [PubMed: 24743020]
43. Nivedhitha G, Anandha Mala GS. Enhanced automatic classification of epilepsy diagnosis using ICD9 and SNOMED-CT. In: Suresh LP, Panigrahi BK, editors. Proceedings of the International Conference on Soft Computing Systems. New Delhi: Springer India; 2016. p. 259–66. (Advances in Intelligent Systems and Computing).
44. Jung H, Lee HY, Yoo S, Hwang H, Baek H. Effectiveness of the use of standardized vocabularies on epilepsy patient cohort generation. Healthc Inform Res. 2022;28(3):240–6. [PubMed: 35982598]

Key Points

- Unstructured data present in EHR are a rich source of medical information; however, their abstraction is labor intensive
- AEP can reduce the need for manual chart review
- Our AEP approach accurately identifies patients with epilepsy from notes, ICD codes, and ASMs
- Neurologists' expertise in identifying keywords and phrases related to diagnoses of epilepsy was vital to creating the AEP models
- The model developed will enable large-scale epilepsy research using EHR data

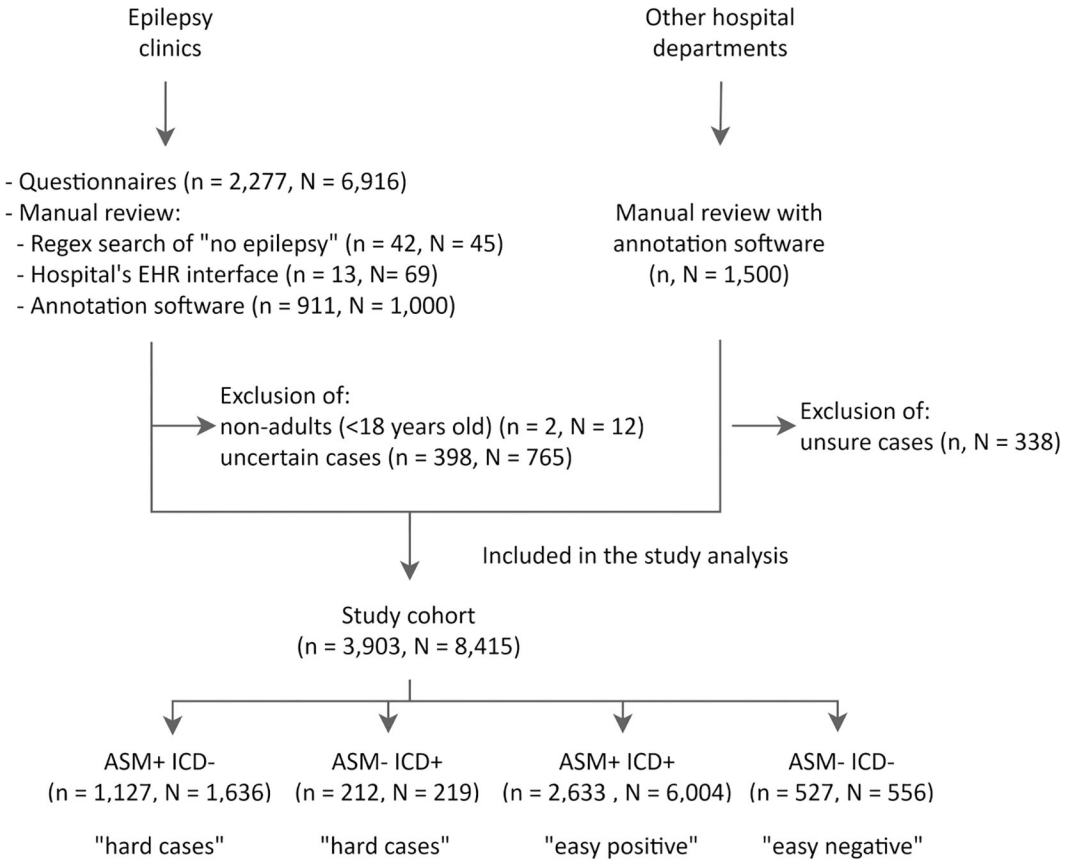

FIGURE 1.

CONSORT diagram for the study cohort, where "*n*" and "*N*" represent the number of patients and visits, respectively. The groups ASM+/− ICD+/− indicate presence/absence of an antiseizure medication (ASM) or epilepsy-related International Classification of Diseases (ICD) code (defined in Subsection 2.3). Patients in the study cohort may have more than one encounter; thus, different encounters for the same patient might be present in each group of hard and easy cases. EHR, electronic health records.

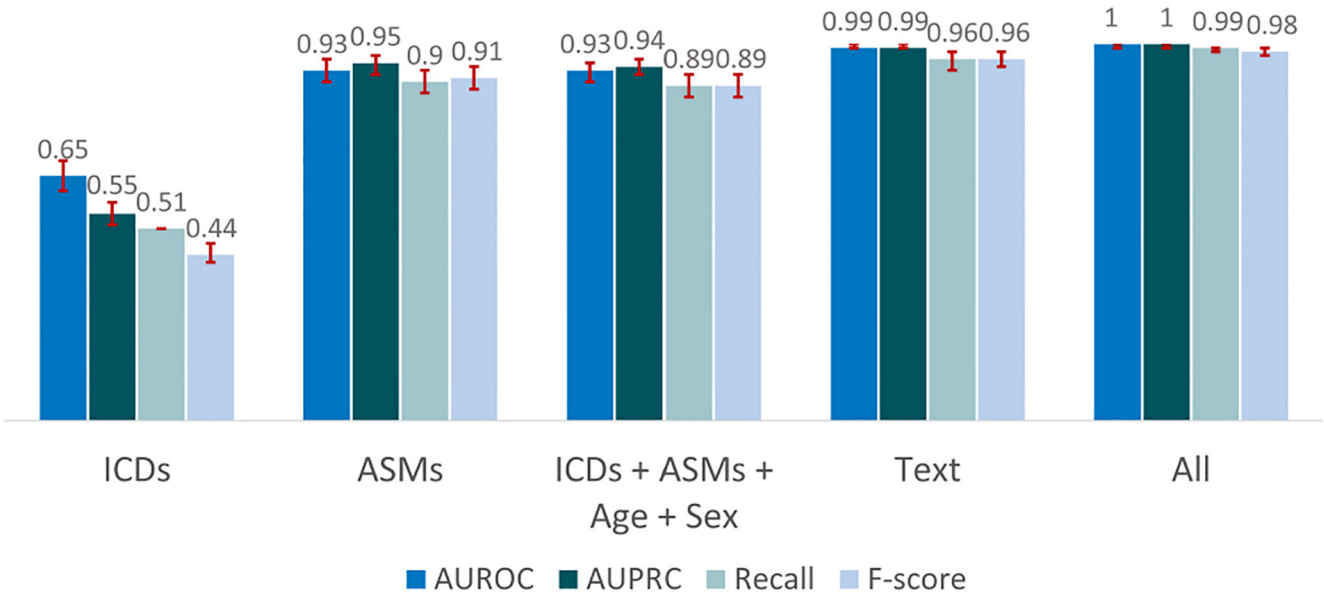

FIGURE 2.

Macro average performance on "hard cases" of the logistic regression classification model for different sets of features in the test set including patients with International Classification of Diseases (ICDs) codes for seizures or antiseizure medications (ASMs). "All" includes text, ICDs, ASMs, age, and sex. AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating characteristic curve.

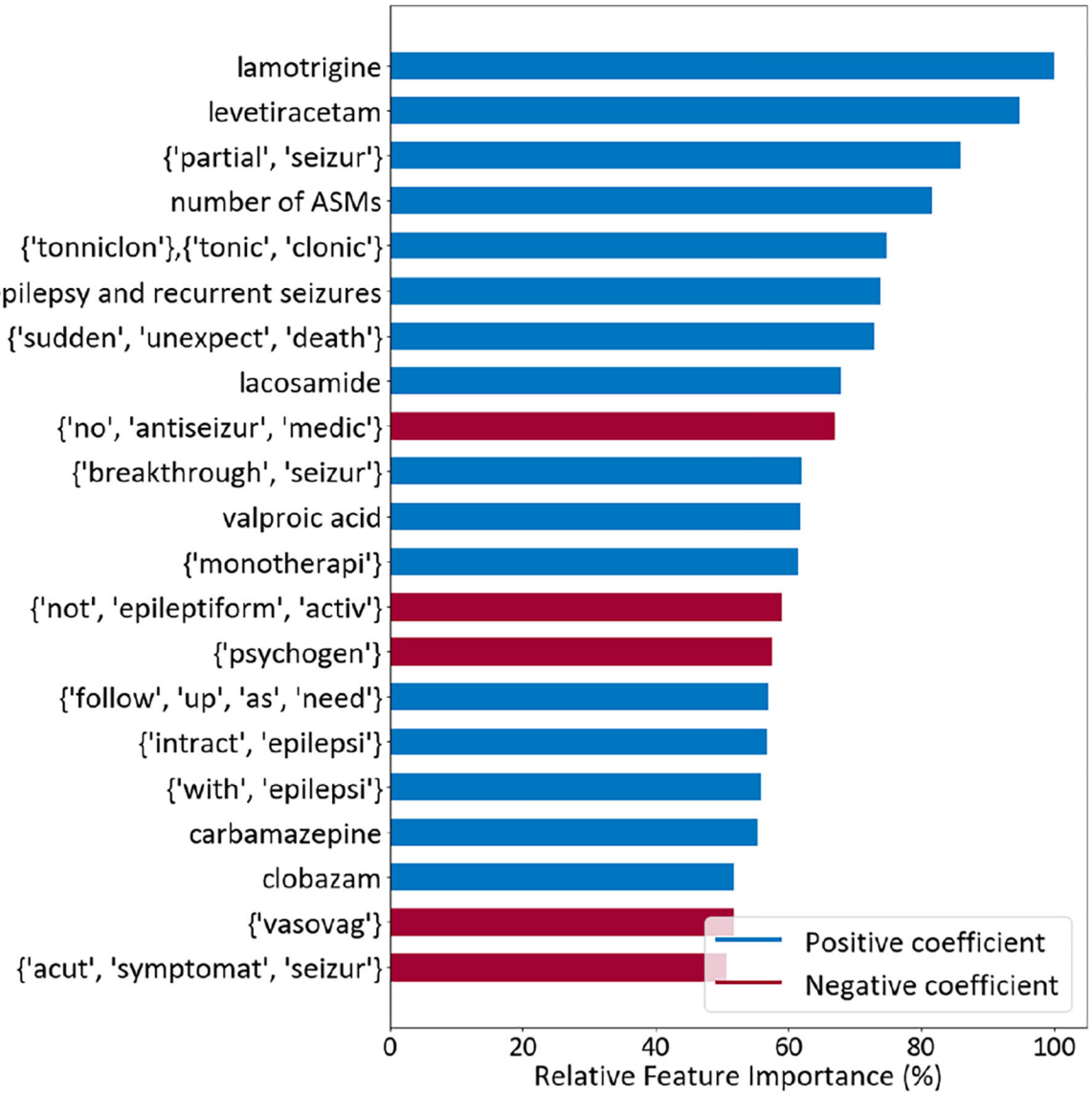

FIGURE 3.

Top 20 most important features of the logistic regression model using the set with all features. For each bag of words in brackets, words are presented in their stemmed form, and they all appear in at least one sentence of the patient visit notes. ASM, antiseizure medication.

TABLE 1

Characteristics of the study cohort.

Characteristic	Study cohort, *n* = 3903
Age, years, mean ± SD[a]	47 ± 18.2
Sex, *n* (%)	
Male	1659 (42.5)
Female	2244 (57.5)
Race, *n* (%)	
Black or African American	217 (5.5)
Other[b]	475 (12.2)
White	3211 (82.3)
Ethnicity, *n* (%)	
Hispanic	263 (6.7)
Unknown	367 (9.4)
Non-Hispanic	3273 (83.9)
Epilepsy diagnosis, *n* (%)	
Positive	2733 (70.0)
Negative	1170 (30.0)
Number of encounters, *N*	8415
Diagnosis, *N* (%)	
Epilepsy and recurrent seizures	5363 (63.7)
Convulsions seizures	1149 (13.7)
Syncope	8 (.1)
Top ASMs, *N* (%)[c]	
Lamotrigine	3134 (37.2)
Levetiracetam	3072 (36.5)
Lorazepam	2131 (25.3)
Lacosamide	1186 (14.1)
Top text features, *N* (%)[d]	
{{“histori”, “seizur”}; {“hx”, ”seizur”}}	6301 (74.9)
{“no”, “seizur”}	5827 (69.2)
{“with”, “epilepsi”}	5785 (68.7)
{{“lamotrigin’}; {“ltg”}}	4505 (53.5)
{“keppra”}	4337 (51.5)
{“focal”}	4274 (50.8)
{“general”, “seizur”}	3968 (47.2)
{{“levetiracetam”}; {“lev”}}	3892 (46.3)
{“febril”}	3877 (46.1)
{“seizur”, “control”}	3177 (37.8)

Note: The number of patients is represented by *n* and the number of visits by *N*.

Abbreviation: ASM, antiseizure medication.

[a] Age at baseline for the first visit in the study period.

[b] "Other" includes unknown, declined, American Indian or Alaska Native, Asian, and Native Hawaiian or other Pacific Islander.

[c] A full list of ASMs is presented in Table S4.

[d] Bag of words for the top 10 text features present in the cohort encounters notes.

TABLE 2

Average performance (95% confidence intervals) for logistic regression model using all features in the test set including patients with International Classification of Diseases codes for seizures or antiseizure medications.

Classes	Model	AUROC	AUPRC	F1 score	Recall	Precision	Specificity
Macro average	LR	1.00	1.00	.98 (.97–.99)	.99 (.98–.99)	.98 (.96–.99)	.99 (.98–.99)
	XGBoost	(.99–1.00)	(.99–1.00)	.96 (.94–.98)	.96 (.94–.98)	.96 (.94–.98)	.96 (.94–.98)
Epilepsy	LR	1.00	1.00	.99 (.98–1.00)	.98 (.97–1.00)	.99 (.99–1.00)	.99 (.97–1.00)
	XGBoost	(.99–1.00)	(.99–1.00)	.98 (.96–.99)	.97 (.95–.99)	.98 (.97–.99)	.96 (.93–.99)
No epilepsy	LR	1.00	.99	.97 (.95–.99)	.99 (.97–1.00)	.96 (.93–.99)	.98 (.97–1.00)
	XGBoost	(.99–1.00)	(.98–1.00)	.94 (.91–.97)	.96 (.93–.99)	.92 (.87–.96)	.97 (.95–.98)

Abbreviations: AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating characteristic curve; LR, logistic regression; XGBoost, extreme gradient boosting model.