Using predictive analytics to identify drug-resistant epilepsy patients

Dursun Delen
Oklahoma State University, USA

Behrooz Davazdahemami
University of Wisconsin, USA

Enes Eryarsoy
Istanbul Sehir University, Turkey

Leman Tomak
Ondokuz Mayis University, Turkey

Abhinav Valluru
Oklahoma State University, USA

Abstract
Epilepsy is one of the most common brain disorders that greatly affects patients’ quality of life and poses serious risks to their health. While the majority of the patients positively respond to the existing anti-epilepsy drugs, others who developed the refractory type of epilepsy show resistance against drug therapy and need to undergo advance treatments such as surgery. Given that identifying such patients is not a straightforward process and requires long courses of trial and error with anti-epilepsy drugs, this study aims at predicting those at-risk patients using clinical and demographic data obtained from electronic medical records. Specifically, the study employs several predictive analytics machine-learning methods, equipped with a novel approach for data balancing, to identify drug-resistant patients using their comorbidities and demographic information along with the initial epilepsy-related diagnosis made by their physician. The promising results we obtained highlight the potential use of machine-learning techniques in facilitating medical decisions and suggest the possibility of extending the proposed approach for developing a clinical decision support system for medical professionals.

Keywords
anti-epileptic drugs, drug resistance, epilepsy, machine learning, predictive analytics, refractory epilepsy

Corresponding author:
Dursun Delen, Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, 700 North Greenwood Avenue, North Hall 302, Tulsa, OK 74106, USA.
Email: dursun.delen@okstate.edu
Introduction

Epilepsy is the third most common chronic brain disorder,¹ which affects approximately 50 million people worldwide (World Health Organization; http://www.who.int/news-room/fact-sheets/detail/epilepsy) with an estimated 2.9 million of them living in the United States (Epilepsy Foundation of America; https://www.epilepsy.com/). It is characterized by unpredictable seizures that differ in type, cause, and severity, and significantly affect the quality of life of individuals and their families in various aspects. Since the introduction of Bromide as an anti-seizure drug, there has been a wide development of therapies effective in decreasing the frequency and severity of seizures in epileptic patients. This class of therapies is referred to as anti-epileptic drugs (AEDs).²

While today AEDs are known to be the most common and effective treatment option for patients with epilepsy (around 60% of patients respond to the first AED prescribed for them), about 15 percent of patients have to spend 2–5 years for finding an effective AED regimen which treats them, and the remaining 25–30 percent tend to be refractory to the AED treatment³⁴ and consequently require to switch to some progressive form of treatment such as surgery or vagus nerve stimulation (VNS).

A report by the Institute of Medicine stresses that although effective treatments are available for many types of epilepsy, timely referrals and access to those treatments are lacking, and epilepsy care and prevention could be enhanced by better data from surveillance and research.⁶ Although there exist some guidelines for treatment of epilepsy, particularly in special cases such as epilepsy in patients with HIV/AIDS⁷ or treatment-resistant patients,⁸ yet in most of the cases, practitioners rely on trial and error for treatment of epilepsy.⁹ This fact makes it difficult to provide patients with proper treatment in a timely manner, whereas treating epilepsy early has a significant positive impact on the patient’s life.

The abundance of healthcare data, resulted from the development of electronic medical records (EMRs) and other similar health information systems in the past decade, presents a great potential to employ Big Data analytics to provide practitioners with clinical decision support systems for making more informed treatment decisions instead of relying on trial and error in case of diseases with unknown causes. For instance, Devinsky et al.⁹ applied Big Data analytics on a data set of medical and pharmacy claims of patients with epilepsy to predict the best AED regimen (for both monotherapy and multi-therapy) that works for each specific patient. In the specific case of epilepsy, another critical decision which has to be taken by the physicians with regard to a given patient is that whether an AED or a non-AED (i.e. surgery or VNS) type of treatment is best effective given the characteristics and background of that patient. In this study, using a large data set of medical records of more than 344,000 epilepsy patients, we train multiple machine-learning classification models with the aim of helping practitioners to identify patients with refractory epilepsy and consequently determine the best type of treatment (i.e. AED vs non-AED) for a given epilepsy patient. Our results suggest that employing Big Data analytics can effectively help physicians in identifying drug-resistant epileptic patients before they undergo multiple stages of AED treatment.

The remainder of the article is organized as follows. In section “Epilepsy types and treatments,” we discuss prior research on epilepsy, its different types, and treatment choices. Section “Materials and methods” presents the data used for conducting the study, followed by the analyses results in section “Results.” We then conclude the article by discussing the results, mentioning theoretical and practical contributions, and providing a summary of our study in sections “Discussion” and “Summary and conclusion.”
Epilepsy types and treatments

Epilepsy is the third most common chronic brain disorder and is characterized by recurrent and unpredictable interruptions of normal brain functions, called epileptic seizures. In fact, epilepsy refers to a group of brain disorders rather than a single disease, and there is still no consensus on a clear definition of that characterizing its causes and mechanisms among scholars or practitioners. Some of the causes identified for epilepsy by the epidemiology researchers include genetic disorder, stroke, brain tumors, traumatic brain injury, and central nervous system infection. Also, various anti-epilepsy drugs have been developed over the past century. However, usually figuring out the main cause of epilepsy in a given patient is not straightforward, which leads physicians to rely on trial and error in treating epileptic patients. Fortunately, around 60 percent of patients with epilepsy positively respond to their first AED, and another 10–15 percent can also be cured using AEDs after a 2–5 year course of trial and errors. Nevertheless, around 30 percent of patients develop refractory (also known as drug resistant) type of epilepsy and neither monotherapies nor combinatory AED regimens significantly treat their condition and they have to undergo surgeries or other progressive treatments such as VNS.

Identifying patients with refractory epilepsy in the early stages of diagnosis is not easy, and it is usually done after multiple unsuccessful courses of AED therapy. During this time, the condition significantly affects the patients’ quality of life and they are at increased risk of sudden unexpected death. Several studies are conducted with the aim of providing guidelines for early identification of refractory epilepsy. For instance, Kwan and Brodie, by doing a prospective study on a group of 525 patients, maintain that patients who had too many seizures before starting AED therapy or those who do not positively respond to the first AED treatment are likely to have refractory epilepsy. Other studies suggest that patients with symptomatic or cryptogenic epilepsy or those with a family history of epilepsy or psychiatric comorbidities (e.g. depression) are least likely to respond to AED therapy.

In addition, by the development of healthcare information systems which has resulted in creation of huge databases of medical transactions, recently, Big Data analytics have been used by some researchers to address sophisticated issues in the health area, including issues regarding epilepsy treatment. Devinsky et al., for example, applied machine-learning techniques to data obtained from a large database of medical claims to train an algorithm for choosing an appropriate AED regimen for individual patients. Another study applied machine learning to the electroencephalogram (EEG) data from 23 pediatric epilepsy patients to detect and classify the onset of seizures in those patients. Other researchers have also addressed more or less the same issues by applying various Big Data analytics algorithms to different types of data. Moreover, recently, deep learning methods have been used for automatic seizure detection in patients with epilepsy.

Though multiple studies have used Big Data methods for addressing issues with regard to the treatment of epilepsy, little research has been conducted to apply Big Data for early identification of patients with refractory epilepsy. This study uses medical transactional data (i.e. historical medication, diagnoses, and procedure records) of patients with epilepsy obtained from an EMR database to train machine-learning classification algorithms and provide a base for a clinical decision support system for early identification of refractory epilepsy.

Materials and methods

To perform this study, we obtained data from the Cerner’s Health Facts data warehouse, one of the most comprehensive EMR databases in the United States containing medical records of more than
Health Informatics Journal 26(1)

63 million unique patients across the country. The data warehouse involves chronological records of individuals’ encounters, diagnosis, medications, lab reports, and procedures along with their demographic information. Figure 1 demonstrates a pictorial depiction of the overall methodology employed in this study.

We used a data set including encounters, medications, diagnosis, and procedure records of 344,473 unique patients who had at least one epilepsy ICD-9 code (i.e. 345.*) in their diagnosis records. Taking advantage of variety in the diagnosis data set which has other major disease indicators for each patient, categorical indicator variables with major diseases were created for each patient. Looking at the descriptive statistics of the patient records shown in Figure 2, each of the eight most frequently occurring diseases were assigned a flag variable to be used as predictors of refractory epilepsy.

In addition, the total comorbidity count was extracted for each patient, as a measure of his or her general wellness, as another predictor. Moreover, we included the patient’s demographic information (i.e. age, gender, race, and marital status) as well as the physician’s initial diagnosis (i.e. the first epilepsy-related ICD-9 diagnosis recorded for each patient) as other potential predictors of refractory epilepsy. The initial diagnosis was included in the model since it potentially carries information about the patient status at the first visit, while it does not involve any trial and error and is purely based on the knowledge of the physicians. Therefore, it possibly can help in determining the real type of epilepsy in the patients. A binary target variable was created to represent whether a patient really had refractory or non-refractory (0) type of epilepsy, based on the latest treatment (i.e. non-AED or AED) recorded for each patient. After data cleaning and integration and dropping patients with too much missing information, we ended up with a data set containing records of 37,024 unique patients, from which 806 (2.2%) had refractory epilepsy and the other 36,218 (97.8%) were considered as patients with a non-refractory type of epilepsy.

The data set was then used to train multiple machine-learning classification algorithms. After putting aside a 30-percent random stratified sample (i.e. 11,108 records) of the whole data set for validating the models, since the training data was severely unbalanced, to avoid any biases in the models, we applied three different balancing approaches to the training data set: (1) the well-known Synthetic Minority Oversampling Technique (SMOTE) was used to balance the data by oversampling the minority group (i.e. patients with refractory epilepsy); (2) a random undersampling approach, which includes all the cases from the minority group along with a random sample of the same size drawn from the majority group cases; (3) an innovative approach involving both undersampling and oversampling techniques. In this approach, we first create a smaller unbalanced data set by including all the patients from the minority group and a random sample drawn from the majority group patients with a size three times the total number of minority patients. The resulted, yet unbalanced, data set then goes through the SMOTE algorithm to be oversampled and become balanced. Table 1 indicates the number of cases in the training data set, resulted from the three balancing approaches.

We then trained multiple classification models using the balanced datasets. Three tree-based machine-learning algorithms, namely, decision tree (DT), random forests (RFs), and gradient boosted trees (GBTs), were employed to build the classification models. Tree-based algorithms are shown to be powerful classifiers, especially in the presence of multiple categorical independent variables. They are also capable of handling missing values, which often exist in transactional health data. While traditional DT has been around for a long time as a decision support tool, RF and GBT are relatively newer machine-learning methods, both falling in the category of ensemble learning algorithms. In the ensemble learning algorithms, the idea is to train a large number of models, as opposed to training only one, and then take the majority vote as the base for decision-making.
Figure 1. A pictorial depiction of the predictive analytics methodology employed in this study.
Specifically, random decision forests, initially proposed by Ho in 1995 and further extended by Breiman in 2001, is an ensemble learning algorithm for supervised classification and regression purposes. DTs trained very deep, despite capturing sophisticated patterns, tend to overfit their training data sets. To avoid this overfitting bias and reduce the variance in them, we can train a multitude (a bag) of deep trees, each trained on a different part (sample) of the same training data set, and then voting/averaging their outcomes to come up with the final decision for each case. This bagging strategy to reduce the bias and variance in deeply trained trees is, in fact, the essence of the RF algorithms. The predictive power of RFs in the healthcare context has been shown in multiple prior studies.

GBTs, however, are ensemble algorithms operating based on the boosting strategy (as opposed to bagging). In this case, instead of training independent trees, trees are trained sequentially, so that each tree learns from the mistakes of the previous one with a specified learning rate to become improved. In fact, boosting refers to this general problem of producing a very accurate prediction rule by combining rough and moderately accurate rules-of-thumb. Boosting algorithms typically work by assigning weights to the observations, putting more weights on the difficult-to-predict cases and less weight on the cases already predicted accurately and, therefore the algorithm continues until it identifies a model that correctly classifies the difficult cases. Multiple prior studies have also shown the potential of GBT for prediction purposes in the healthcare area.

Results

Among various machine-learning algorithms used to make classification models, three tree-based algorithms, namely, DT, RFs, and GBT, provided the best results in terms of classification accuracy.
It was consistent with our expectations since tree-based algorithms are particularly more powerful than algorithms such as artificial neural networks (ANN) or support vector machine (SVM) when there are several categorical independent variables involved. Table 2 presents the accuracy statistics of the models on the validation data, using training data sets resulted from oversampling, undersampling, and the combinatory approach. For each model and using each training data set, the table demonstrates sensitivity (true positive rate), specificity (true negative rate), F-measure, overall accuracy, and the area under the receiver operating characteristic (AUROC) curve.

As shown in Table 2, RF and GBT, the two ensemble classification algorithms, provided generally more accurate results than the simple DT model, with GBT providing the best results with regard to all the three training data sets.

In addition, looking into the results of the three GBT models (i.e. the last column of Table 2), it looks like the training data set involving the combinatory balancing approach (bold values in Table 2) provided better results than both undersampling and oversampling approaches. Particularly, the model trained using the combinatory approach balanced data set clearly outperforms the other two models in terms of both F-measure and AUROC, whereas it performs almost as equally well as the other two models in terms of sensitivity, specificity, and overall accuracy. The F-measure statistics represents a weighted harmonic mean of sensitivity and specificity of the model. It is particularly crucial in testing models using unbalanced data sets, especially where one class is more important than the other. In this case, since correctly predicting a refractory epilepsy case is more important to us than correctly labeling a case as non-refractory, the classifier with a higher F-measure is preferred. Also, the higher AUROC for this model suggests its relatively higher power in distinguishing between refractory and non-refractory epilepsy cases in comparison with the other alternatives.

According to the results, this best model correctly predicted the existence of refractory epilepsy in 74.9 percent of patients who actually had that type of epilepsy. It suggests that given historical information about patients’ comorbidities as well as their demographic information and the initial diagnosis of the physician, our best model was able to predict 181 out of 242 actual refractory epilepsy cases that could take 2–5 years, on average, to be diagnosed after multiple trial and errors.

### Table 2. Accuracy measures of the classification models.

<table>
<thead>
<tr>
<th>Model type</th>
<th>Measure type</th>
<th>DT (%)</th>
<th>RF (%)</th>
<th>GBT (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oversampling</td>
<td>Sensitivity</td>
<td>52.90</td>
<td>37.20</td>
<td>79.80</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>77.90</td>
<td>88.90</td>
<td>70.10</td>
</tr>
<tr>
<td></td>
<td>F-measure</td>
<td>12.10</td>
<td>12.40</td>
<td>13.40</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>73.70</td>
<td>87.60</td>
<td>79.80</td>
</tr>
<tr>
<td></td>
<td>AUROC curve</td>
<td>0.777</td>
<td>0.789</td>
<td>0.823</td>
</tr>
<tr>
<td>Undersampling</td>
<td>Sensitivity</td>
<td>54.90</td>
<td>76.90</td>
<td>74.90</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>75.10</td>
<td>71.30</td>
<td>73.80</td>
</tr>
<tr>
<td></td>
<td>F-measure</td>
<td>13.30</td>
<td>12.40</td>
<td>11.50</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>73.90</td>
<td>77.20</td>
<td>74.20</td>
</tr>
<tr>
<td></td>
<td>AUROC curve</td>
<td>0.778</td>
<td>0.814</td>
<td>0.814</td>
</tr>
<tr>
<td>Under + oversampling</td>
<td>Sensitivity</td>
<td>55.80</td>
<td>78.90</td>
<td>74.90</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>76.40</td>
<td>64.70</td>
<td>75.40</td>
</tr>
<tr>
<td></td>
<td>F-measure</td>
<td>48.30</td>
<td>55.80</td>
<td>59.30</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>71.20</td>
<td>73.10</td>
<td>75.30</td>
</tr>
<tr>
<td></td>
<td>AUROC curve</td>
<td>0.793</td>
<td>0.804</td>
<td>0.829</td>
</tr>
</tbody>
</table>

DT: decision tree; RF: random forest; GBT: gradient boosted tree; AUROC: area under the receiver operating characteristic.
We further compared the model predictions against the initial physicians’ diagnosis as well as the actual type of epilepsy in the patients. Interestingly, it was revealed that with regard to 148 patients who actually developed refractory epilepsy (ICD-9: 345.91), while the initial diagnosis of the physician was other types of epilepsy, our model correctly predicted that those patients are resistant to AED treatment. In addition, in the case of 14 patients for whom the initial diagnosis of the physician was refractory epilepsy, our model correctly classified them as patients with non-refractory epilepsy. There was only one patient for whom the initial diagnosis of the physician was correctly refractory epilepsy, whereas our model incorrectly labeled him or her with a non-refractory type of the disorder. In addition, the initial refractory diagnoses of physicians were correct in only 23 out of 242 actual cases (i.e. 9.5%), compared to 181 (i.e. 74.9%) by our classification model. Overall, these results suggest the superior performance of our classification model in making initial diagnoses compared to the physicians.

Having the superior classification model identified, we further investigated how each of the predictors contributed to the model sensitivity. To this end, we dropped predictor variables one at a time from our data and ran the best classification model with the same settings. Each time we recorded the model’s sensitivity (i.e. model’s ability to correctly classify actual refractory epilepsy cases) to be compared to that of the original model. In each iteration, we calculated the difference between the original model sensitivity and the sensitivity resulted from dropping each corresponding variable. The max–min normalized sensitivity difference values were used as a measure of variables’ relative importance as depicted in Figure 3. As shown in this chart, comorbidity count as a measure of overall wellness turned out to be the most important variable in correctly predicting refractory epilepsy, followed by drug abuse, hypertension, and hyperlipidemia disorders.

Interestingly, the chart indicates that the variable FIRST_EP_DIAG representing the initial physician diagnosis is the second least important predictor in determining the type of epilepsy. In fact,
the results provide support for the idea that physicians’ knowledge for understanding the causes and diagnosing the drug resistance of epilepsy cases is yet far from perfection and relies on trial and errors.

Discussion

Using historical medical records of more than 34,000 unique patients with epilepsy, this study provides a machine-learning model to identify patients with the refractory type of epilepsy (i.e. patients’ resistant to the medication treatment) before they go through multiple courses of medical trial and error. The proposed model was able to correctly classify 75 percent of patients who actually had developed refractory epilepsy using their comorbidities, demographic information, and the initial diagnoses made by their physicians.

From a medical viewpoint, this study contributes by providing a base for designing a clinical decision support system to help physicians identify drug-resistant epilepsy patients more accurately and in a shorter time than it usually takes. Generally, the comparisons made between the model predictions and the initial physician diagnoses highlight the considerable potential change that employing such a decision support system may result in. While a patient classified by the proposed model as a drug-resistant patient may still respond to AED treatments, we believe that knowledge about his or her high likelihood of having refractory epilepsy (provided by the model) can considerably decrease the number of trial and errors and lead to lowering risks to his or her life due to long AED treatment courses.

In terms of methodology, this study provides a novel innovative approach for balancing severely unbalanced training data by combining traditional oversampling and undersampling techniques, which was shown to be more effective than both of them in terms of model accuracy statistics. Future research may employ this approach in other contexts for unbalanced datasets to validate its power in providing more accurate classifications.

In addition, our results provide additional evidence for the higher predictive power of ensemble machine-learning methods (i.e. RF and GBT) in comparison with traditional algorithms in capturing sophisticated patterns in the data. While statistical and regular machine-learning techniques train a single model (either linear or non-linear) to reflect the relationship between the variables, ensemble algorithms sample the data hundreds of times and use those samples to build hundreds of classification models. Then to classify a new case, they vote from the created models to specify the final class. This way, instead of a single model, which is subjected to sample randomization errors, many models are employed to yield classifications.32

Although this research provides promising results for identifying patients with refractory epilepsy, it was limited to using only demographics and comorbidity features of the patients for classifying them, future research may extend this study by including more features such as lab tests, surgery records, and/or EEG results, which are previously shown to be useful for epilepsy diagnoses.37,38

Summary and conclusion

Epilepsy, as the third most common chronic brain disorder, poses many risks to the lives of millions of people all around the world. While AED treatments are proven to be an effective way to cure this disorder in the majority of the patients, yet there are around 30 percent of the patients who do not respond to AED treatments (known as patients with refractory epilepsy) and require progressive treatments. Identification of such patients is not straightforward though and require many long courses of trial and error with AEDs by the physicians, whereas the patients’ quality of life is highly affected and they are at the risk of sudden death during that period.
This study employs historical comorbidities and demographic information of the patients along with the initial diagnosis made by their physicians at the first epilepsy-related visit to predict whether or not each patient developed refractory epilepsy (i.e. would resist to AED treatments).

Our promising results, especially compared to the initial diagnoses of the physicians, highlight the power of analytics in solving practical health problems. The proposed approach can be extended and used as a clinical decision support system to aid the physicians in identifying drug-resistant epilepsy patients in a shorter time.

Author’s note
Behrooz Davazdahemami is now affiliated with University of Wisconsin - Whitewater, USA.

Acknowledgements
This work was conducted with the data obtained from the Cerner Corporation’s Health Facts data warehouse of EMRs, provided by Oklahoma State University (OSU), Center for Health Systems Innovation (CHSI). Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Cerner Corporation, OSU, or CHSI. The authors would also like to acknowledge Ms Elvena Fong, health data analytics program manager at CHSI, for her support in extracting and understanding of the health care data.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Behrooz Davazdahemami https://orcid.org/0000-0003-2885-6014

References


