HHS Public Access

Author manuscript

Epilepsy Behav. Author manuscript; available in PMC 2019 December 01.

Published in final edited form as:

Epilepsy Behav. 2018 December ; 89: 118–125. doi:10.1016/j.yebeh.2018.10.013.

Early Prediction of Drug Resistant Epilepsy using Claims Data

Sungtae An, MSc[1], **Kunal Malhotra, PhD**[1], **Cynthia Dilley**[2], **Edward Han-Burgess**[2], **Joseph Robertson**[2], **Chris Clark**[2], **M. Brandon Westover, MD, PhD**[*,3], and **Jimeng Sun, PhD**[*,1]

[1]Georgia Institute of Technology, College of Computing, Atlanta, GA, USA

[2]UCB Pharma, Smyrna, GA, USA

[3]Massachusetts General Hospital, Department of Neurology, Boston, MA, USA

Abstract

Objective: To determine whether information in medical and pharmacy claims data can predict, at the time of prescribing the first antiepileptic drug (AED), which patients with epilepsy will become resistant to AEDs.

Method: We analyzed longitudinal claims data from 1,376,756 patients with epilepsy from 2006 to 2015. Of these, 582,258 satisfied all inclusion criteria; 49,916 were ultimately AED resistant, operationally defined as a patient with claims filed for at least 4 distinct AEDs. We constructed 1,270 candidate predictors ("features") reflecting demographics, comorbidities, medications, procedures, epilepsy status, and payer status to characterize the cohort. On the training dataset (528,640 patients) we performed ANOVA F-value tests to select predictive features and trained several prediction algorithms, including logistic regression, support vector machines (SVM), and random forests. A model with only age and gender was used as a benchmark model.

Results: On a held-out test set (53,618 patients), the best model achieves an area under the receiver operating characteristic (ROC) curve (AUC) [95% CI] of 0.753 [0.747, 0.759], compared to 0.664 [0.658, 0.671] for the benchmark model. Moreover, predicted probabilities for drug resistance closely match the observed frequencies. Compared to waiting for 2 AED failures, our model predicts drug resistance on average 2.25 years earlier.

Corresponding Author: Jimeng Sun, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30313, Phone 404-894-0482, jsun@cc.gatech.edu.

*Co-senior authors.

Author Contributions:

Sungtae An, analysis and interpretation of data, drafting the manuscript

Kunal Malhotra, analysis and interpretation of data, drafting the manuscript

Jimeng Sun, study concept and design, study supervision

M. Brandon Westover, study design and result interpretation

Cynthia Dilley, study design and result interpretation

Chris Clark, initial study design and conceptualization of the project

Joseph Robertson, acquisition and management of data and project management

All authors reviewed and revised the manuscript.

Author Disclosures:

Sungtae An – None

Kunal Malhotra – None

Jimeng Sun – Receives funding from NSF-IIS-1418511 and CCF-1533768, Children's Healthcare of Atlanta, UCB

M. Brandon Westover – Receives funding from NIH-NINDS (1K23NS090900) and consults for UCB.

Cynthia Dilley, Chris Clark, and Joseph Robertson – are employees of UCB

Study Funded by UCB

Conclusion: Predictive models created from large claims data using machine learning methods can accurately predict which patients with epilepsy will prove drug resistant at the time of prescribing the first AED. The ability to predict refractoriness may help patients consider alternative therapies earlier in the course of their epilepsy.

Keywords

Drug resistant epilepsy; refractory epilepsy; big data; predictive analytics; machine learning

INTRODUCTION

Patients with epilepsy (PWE) suffer decreased quality of life, productivity, life expectancy, and chronic medication side effects. Approximately 50% of PWE achieve seizure control with the first anti-epileptic drug (AED) prescribed[1]. Over the next 2–5 years another 13% become seizure-free with a second AED, and another 4% after trying 3 or more drugs[1]. The International League Against Epilepsy (ILAE) has thus recently defined drug resistant epilepsy (DRE) as "*failure of an adequate trial of two tolerated, appropriately chosen and used AED schedules (whether as monotherapies or in combination) to achieve sustained seizure freedom.*"[2]

Recognition of DRE is often delayed, evidenced by the observation that physicians often continue AED trials for more than 16 years before referring DRE patients for epilepsy surgical evaluation[3],[4]. Personalized tools to predict an individual's probability of drug resistance might allow patients and physicians to consider alternative interventions (e.g. surgery, or drug trials) earlier.

Comprehensive medical and pharmacy claims provide a unique data source, which has not been studied in the literature for DRE prediction. Despite the noisy nature of claims data, big data approaches can often overcome quality limitations[5, 6, 7, 8, 9]. In this work we study a large cohort of PWE (n=1,376,756) with a follow-up time of 10 years (2006 to 2015) and investigate whether machine learning models can predict the probability that a patient will have DRE at the time when they are prescribed the first AED.

METHODS

Comprehensive claims data including diagnosis, procedures and pharmacy claims spanning a period of 10 years ranging from January 2006 through December 2015 were collected from different regions of the United States by Intercontinental Marketing Services (IMS) Health. Table 1 summarizes cohort characteristics including medications used to identify patients in the database determined to have epilepsy (see below for case definition). The data we used was "open source": it was not derived from a specific payer but rather was aggregated from many different sources. The benefit of this data is that it is payer agnostic. Thus, patients who do not stay with the same job/employer for 10 years can still be represented in an open data set but not in a closed data set. Open source data is more representative of the average patient population from both a stability perspective and a payer variation perspective.

Our analysis considers 23 AEDs. We classified three as rescue medications rather than daily AEDs (diazepam, clonazepam, and lorazepam). While clonazepam is also used as a chronic therapy, about 69%, of uses in the data appear to be short term which more closely resembles rescue therapy. Table e-1 shows the complete list of AEDs used in the study.

Data Preparation

Using claims data to model DRE requires making some informed assumptions. In principle changes in AED regimens can occur for multiple reasons besides inefficacy, including drug-related side effects or changes in insurance coverage. With claims data we can only observe discontinuations of medications and additions of new medications. Therefore, we treat all AED changes satisfying certain conditions (see below) as drug failures.

The AED(s) being taken at any given time, the times at which AEDs were changed, and the patient's final state (drug resistant: having tried 4 or more distinct AEDs, vs not drug-resistant: having tried only 1 AED without change for more than 1 year) were inferred from the claims dataset as described in the Supplemental Material and in the definitions outlined in the following subsections.

Cohort Definition

Inclusion criteria.—Patients from the IMS claims dataset to be included in our study had to satisfy the following criteria:

1. *Epilepsy diagnostic criteria*:
    a. The patient has at least one epilepsy diagnosis claim with ICD9 code 345.* or an equivalent ICD10 code G40.*, or 2 claims for convulsions, including ICD 9 code 780.39 or ICD10 code R56.9.
    b. The patient has at least one claim for an AED medication.
    c. The first AED is a monotherapy with at least 60 days supply.
2. *Age*: Patients are at least 16 years old at the time of the first AED prescription.
3. *Data completeness*: Patients have active pharmacy claims every quarter of the year.
4. *Drug resistance:* Patients qualify as having either DRE or non-drug resistant epilepsy (non-DRE) according to the study definitions of AED failure and DRE later in this section.

Criteria 1a and 1b use either a formal epilepsy code OR a convulsion code plus an AED prescription as the proxy to identify patients with epilepsy, following prior literature[10]. Criterion 1c helps ensure that the AED is intended to be a long term treatment rather than an interim prescription until the patient can see another physician (e.g. an epileptologist). Criterion 3 helps avoid including patients who appear to be drug resistant simply because of non-compliance with AEDs due to non-reporting. Criterion 4 ensures clean distinction of DRE from non-DRE. Table 2 summarizes the impact of inclusion and exclusion criteria on the cohort size.

Time of prediction (index date), observation window, and evaluation window

We call the time at which we wish to make predictions the “index date”. This is the time at which we predict whether the patient will ultimately have DRE. We define the index date as the time of prescription of the first AED. The index date is identified from the IMS claims dataset by scanning longitudinal records until reaching the first record of a monotherapy AED.

The index date defines a dividing point in a patient’s medical history timeline. The period before the index date is the *observation window,* and the period after becomes the *evaluation window*. The evaluation window begins immediately after the index date and extends to the patient’s last available record. The prediction made at the index date must be based exclusively on data available leading up to and including the index date. Data from after the index date is not used in making predictions. Rather, this data is used exclusively to judge whether drug resistance predictions turn out to be correct. Figure e-2 illustrates the observation and evaluation windows. The index date is set equal to the time of first AED prescription.

Definition of AED Success and Failure

Treatment success or failure is not directly available from claims data. As a proxy we consider an AED regimen a success if the patient consistently filled the same AED(s) over at least 1 year without the addition of another AED or cessation of AED therapy. Likewise, we operationally define an AED failure to occur whenever a different AED is filled, either as a replacement of the current AED or as an addition to the current AED.

Definition of drug resistant epilepsy (DRE): We aim to predict, at an early stage, which patients will develop DRE. Because seizure status (seizure free vs having ongoing seizures) is not directly captured in the claims dataset, we use the number of distinct AED regimens prescribed over time as a proxy indicator for DRE[11, 12]. We operationally define drug resistant and non-drug resistant cases as follows:

Drug resistant epilepsy (DRE): Patients are considered drug resistant when they fail 3 or more AEDs.

Non-drug resistant epilepsy (non-DRE): Patients are considered non-drug resistant if they remain on a single AED for at least one year. That is, non-drug resistant patients remain on the first AED prescribed one or more years.

These definitions create a relatively clean distinction between the two groups of interest. Patients who fail only one or two AEDs during the time covered in the IMS claims dataset are considered to be in an indeterminate state and are excluded from the analysis.

Feature Engineering

Our modeling approach selects predictive features from a set of 1,270 candidate features extracted from each patient’s observation window. We do not include information about the first AED, because the model is intended for use at the time when a physician is selecting

the first AED. The observation window includes the period before and including the index date. These features are categorized into 5 types: *Demographics, Comorbidities, Insurance Policy, Treatments,* and *Encounters.* These features are further described in the Supplementary Material.

We compare the performance of several classification algorithms, including Linear Support Vector Machines (SVM), Multivariate Logistic Regression, and Random Forests. Model outputs are post-processed during model training to improve calibration using isotonic regression[13] with 5-fold cross-validation.

Statistical analysis of the quality of predictions made by the model

We test the final model on the held out test set. Testing is performed by feeding to the predictive models the data from the observation window and having them predict whether the patient becomes drug resistant during the evaluation window (Figure e-2). We measure Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC), referred to as AUC, and the Average Precision (AP), equivalent to the area under the Precision-Recall (PR) curve, to evaluate model performance. We evaluate 95% confidence interval (CIs) by 10,000 rounds of bootstrapping.

To assess the added value of the engineered features and model optimization procedures described above, we compare our best performing models with a baseline model that makes predictions based only on age and gender. To gain qualitatively insights into which features contribute most to predictive accuracy, we calculate the information gain[14] of features in the Random Forest model, and perform a qualitative analysis of the top 20 features with largest information gain.

Model calibration (how well model predictions match the observed frequencies of DRE) are assessed by calculating Breier scores and calibration plots.

RESULTS

Patient cohort

Among 1,376,756 patients in the IMS claims dataset, 582,258 meet the study inclusion criteria. Among those, 49,916 fail 3 or more AEDs during the evaluation window and are thus categorized as having developed DRE; 532,342 patients continue on the same AED prescribed at the index date throughout the entire evaluation window and are classified as non-DRE during the observation period. A subset of patients is designated as the *training dataset* and used to select an optimal set of predictive features and to tune the model parameters; overall we refer to this process as *training the model*. The rest of the data is assigned as the *testing dataset.* It is used to assess the predictive power of the trained model and is never used for training or tuning the model. Table 3 and Supplemental Material describe these datasets in detail.

Univariate analysis

Univariate statistics characterizing the case and control group, including age, gender, and key epilepsy-related comorbidities, are presented in Table 4.

Model performance: binary predictions

The highest AUC [95% CI] achieved by each experimental model is 0.753 [0.747, 0.759] for Random Forest, compared with 0.732 [0.725, 0.738] for Logistic Regression, and 0.720 [0.714, 0.726] for SVM, by using 508, the top 40%, features for all models. We conducted experiments with varying numbers of features, from 1% to 100% of 1270 candidate features, chosen by an ANOVA F-value test[15] which provides a univariate measure of dependence between target values and each feature. Figure e-3 shows the performance by different number of features.

With the baseline model of age and gender features, the classifiers obtain an AUC of 0.664 [0.658, 0.671]. Figure 1A shows the area under the ROC curve for the experimental and the baseline models.

The highest AP achieved by each model is 0.325 [0.314, 0.336] for Random Forest, 0.315 [0.304, 0.326] for Logistic Regression, and 0.288 [0.278, 0.299] for SVM. Figure 1B compares the PR curves. Our predictive models, especially Random Forest, substantially outperform the baseline model in terms of both AUC and AP.

To further assess the quality of the prediction models developed in this study, we show calibration reliability plots for SVM, Random Forest and Logistic Regression in Figure 2. The Brier score is calculated by the mean squared difference between the observed frequency and predicted probability or DRE[16]. The diagonal in the plot indicates the performance that perfectly calibrated predictions would have, in which the predicted probabilities of drug resistance precisely match the observed percentages. A well calibrated model would have the curve as close as possible to this diagonal. Calibration scores for our predictive models, whose Brier scores [95% CI] are 0.102 [0.107, 0.111] for Random Forest, 0.103 [0.101, 0.105] for Logistic Regression, and 0.104 [0.102, 0.107] for SVM, are substantially better than the baseline model, with Brier score of 0.109 [0.107, 0.111].

Qualitative Results

We next investigate the predictive power of each feature selected by the model training procedure for inclusion in the final predictive model. The top 20 features with highest information gain values are listed in the Table 5. Predictive models need not have features that are readily interpretable. Nevertheless, we are able to group the top 20 features into three relatively natural groups, including indicators of a patient's level of activity within the medical system (e.g. frequency of tests, procedures, and medical visits); indicators of comorbidities (e.g. age, need for drugs associated with cerebrovascular disease risk); and indicators of the complexity and intensity of a patient's epilepsy (e.g. frequency of visits associated with an epilepsy diagnosis code). These three groups seem relatively similar to those found in prior analyses of DRE risk factors using non-claims data.

We also analyze the amount of potential time saved by our model relative to status quo of predicting refractoriness after a patient fails their second AED treatment. Figure 3 shows the survival analysis (Nelson-Aalen) estimate[17, 18] of cumulative distribution fraction of patients who have failed two AEDs on our test dataset. The time axis measures the period between the first AED prescription (the index date) and the third AED prescription (the second

failure); control cases are treated as censored at the time they drop out of the dataset, since we have not observed their third AED prescription by the definition. The mean and median value of time taken to be determined as DRE patient in this way are 2.26 and 1.86 years respectively. Consequently, our method predicts DRE on average 2 years 3 months before a patient fails 2 AEDs, the usual time needed to be identified as having DRE[2]. Given the longitudinal time constraint of needing to have failed at least 2 drugs within the observation period, real world clinical data may reflect much longer times to failure.

Discussion

Seizures prove resistant to currently available AEDs in approximately one third of patients with epilepsy[19]. Patients with drug resistant epilepsy (DRE) suffer for years through repeated failed AED trials before being referred for trials of new drugs or epilepsy surgery, in part because accurate models to predict drug resistance have been lacking. Herein we have describe machine learning models that use information latent in large claims data to accurately predict whether a patient with epilepsy will become drug resistant (AUC 0.753 [95% CI 0.747, 0.759]). Our model provides predictions that are well-calibrated, in that probabilities assigned by the model are in good agreement with the observed proportions of patients who ultimately become drug resistant (Brier score 0.102 [95% CI 0.100, 0.104]). Our model is designed to be used at the time of the first AED prescription, i.e. at the time when epilepsy is generally first diagnosed. Our results suggest that this early prediction capability has the potential to identify patients at high risk for DRE 2.25 years earlier than the current practice of waiting for a patient to fail two AEDs before diagnosing DRE[2].

Several prior studies have addressed drug resistance in epilepsy. However, most have provided univariate analysis of clinical findings from relatively small cohorts rather than true predictive models[20, 21]. By contrast, we have constructed predictive models which combine multiple predictive factors latent in claims data. Our predictive models are made possible by the availability of “Big Data”, comprising claims from 582,258 patients, and by recent advances in machine learning and computing.

A strength of our study is the use of a claims data set that was “open”[22], which means the patients are linked across multiple payers over time. We chose an open dataset to approximate the distribution of patient demographics, comorbidities, and payers and payer types across United States. The use of an open dataset also allows tracking patients over multiple years regardless of changes in insurance and employment status. Nonetheless, working with open data sets requires care due to variation in reporting across different pharmacies.

While the large size and statistical power of our dataset is a strength of our study, all claims data studies are subject to limitations[23, 24]. First, identification of patients with epilepsy cases based on diagnostic codes and AED prescriptions does not always produce results identical with those found on direct chart review[23]. Second, identification of drug patients with seizure freedom vs drug resistance by changes in AED use is indirect. Nevertheless, while treatment changes may arise from several causes (e.g. inefficacy, side effects, noncompliance), we believe that these probably correspond reasonably well with AED failures reported in other studies of drug resistance. Similarly, while treatment stability can

be reasonably assumed to signify good seizure control, it is possible that in some cases this instead simply represents physician or patient resistance to make AED changes.

Future work should investigate how well our models perform against the gold standard of drug resistance based on direct retrospective review of medical notes or in prospective clinical studies. Another promising albeit more challenging direction is to develop predictive models that extract information directly from the electronic health record (EHR). However, EHR data can be challenging due to both the infrequency of observations as well as missing data from visits to other health care providers which may not be incorporated in some EHR systems. Also, the nascency of electronic systems can challenge the need for lengthy longitudinal data with robust samples of 10 years or more.

CONCLUSION

Predictive models based on large claims data can accurately predict DRE at the time of first AED prescription. Our model accurately predicts drug resistance on average 2.25 years earlier than the standard requirement of failing 2 AEDs before predicting drug resistance. These results show how predictive modeling and “Big Data” may help drug resistant patients consider alternative therapies earlier in the course of their epilepsy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Kwan P, Brodie MJ. Early Identification of Refractory Epilepsy. N Engl J Med 2000;342:314–319. [PubMed: 10660394]
2. Kwan P, Arzimanoglou A, Berg AT, et al. Definition of drug resistant epilepsy: consensus proposal by the ad hoc Task Force of the ILAE Commission on Therapeutic Strategies. Epilepsia 2010;51:1069–1077. [PubMed: 19889013]
3. Engel J Jr, Wiebe S, French J, et al. Practice parameter: temporal lobe and localized neocortical resections for epilepsy: report of the Quality Standards Subcommittee of the American Academy of Neurology, in association with the American Epilepsy Society and the American Association of Neurological Surgeons. Neurology 2003;60:538–547. [PubMed: 12601090]
4. Haneef Z, Stern J, Dewar S, Engel J Jr., Referral pattern for epilepsy surgery after evidence-based recommendations: a retrospective study. Neurology 2010;75:699–704. [PubMed: 20733145]
5. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 2015;115:211–252.
6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016: 770–778.
7. Bahdanau D, Chorowski J, Serdyuk D, Brakel P, Bengio Y. End-to-end attention-based large vocabulary speech recognition. Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on; 2016: IEEE: 4945–4949.
8. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–2410. [PubMed: 27898976]
9. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–118. [PubMed: 28117445]
10. Faught E, Richman J, Martin R, et al. Incidence and prevalence of epilepsy among older US Medicare beneficiaries. Neurology 2012;78:448–453. [PubMed: 22262750]

11. Devinsky O Patients with Refractory Seizures. New England Journal of Medicine 1999;340:1565–1570. [PubMed: 10332020]
12. Berg AT, Kelly MM. Defining intractability: comparisons among published definitions. Epilepsia 2006;47:431–436. [PubMed: 16499772]
13. Chakravarti N Isotonic median regression: a linear programming approach. Mathematics of operations research 1989;14:303–308.
14. Michalski RS, Carbonell JG, Mitchell TM. Machine learning: An artificial intelligence approach: Springer Science & Business Media, 2013.
15. Lowry R. Concepts and applications of inferential statistics. 2014.
16. Brier GW. Verification of forecasts expressed in terms of probability. Monthly weather review 1950;78:1–3.
17. Nelson W. Theory and applications of hazard plotting for censored failure data. Technometrics 1972;14:945–966.
18. Aalen O Nonparametric inference for a family of counting processes. The Annals of Statistics 1978:701–726.
19. Epilepsy Foundation. Epilepsy Statistics [online]. Available at: http://www.epilepsy.com/ Accessed 21 Feb.
20. Viteva E. Predictors and Typical Clinical Findings of Refractory Epilepsy. American Journal of Clinical Medicine Research 2014;2:26–31.
21. Semah F, Ryvlin P. Can we predict refractory epilepsy at the time of diagnosis? Epileptic disorders 2005;7:10–13.
22. Kee VR, Gilchrist B, Granner MA, Sarrazin NR, Carnahan RM. A systematic review of validated methods for identifying seizures, convulsions, or epilepsy using administrative and claims data. Pharmacoepidemiology and drug safety 2012;21:183–193. [PubMed: 22262605]
23. 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:n/a-n/a.
24. Devinsky O, Dilley C, Ozery-Flato M, et al. Changing the approach to treatment choice in epilepsy using big data. Epilepsy & Behavior 2016;56:32–37. [PubMed: 26827299]

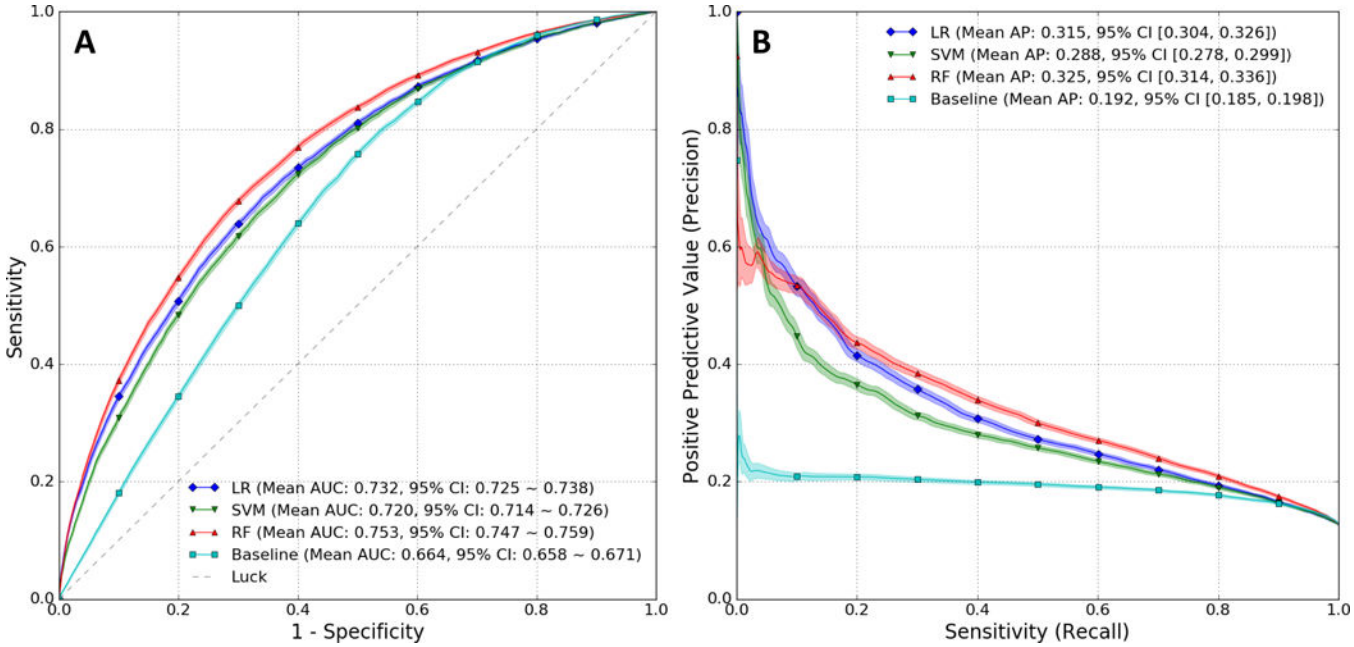


Figure 1. (a) ROC curves (b) PR curves

Both ROC and PR curves are generated from 10,000 times of bootstrapped results. Solid lines are mean curves and shaded regions are their 95% confidence intervals.

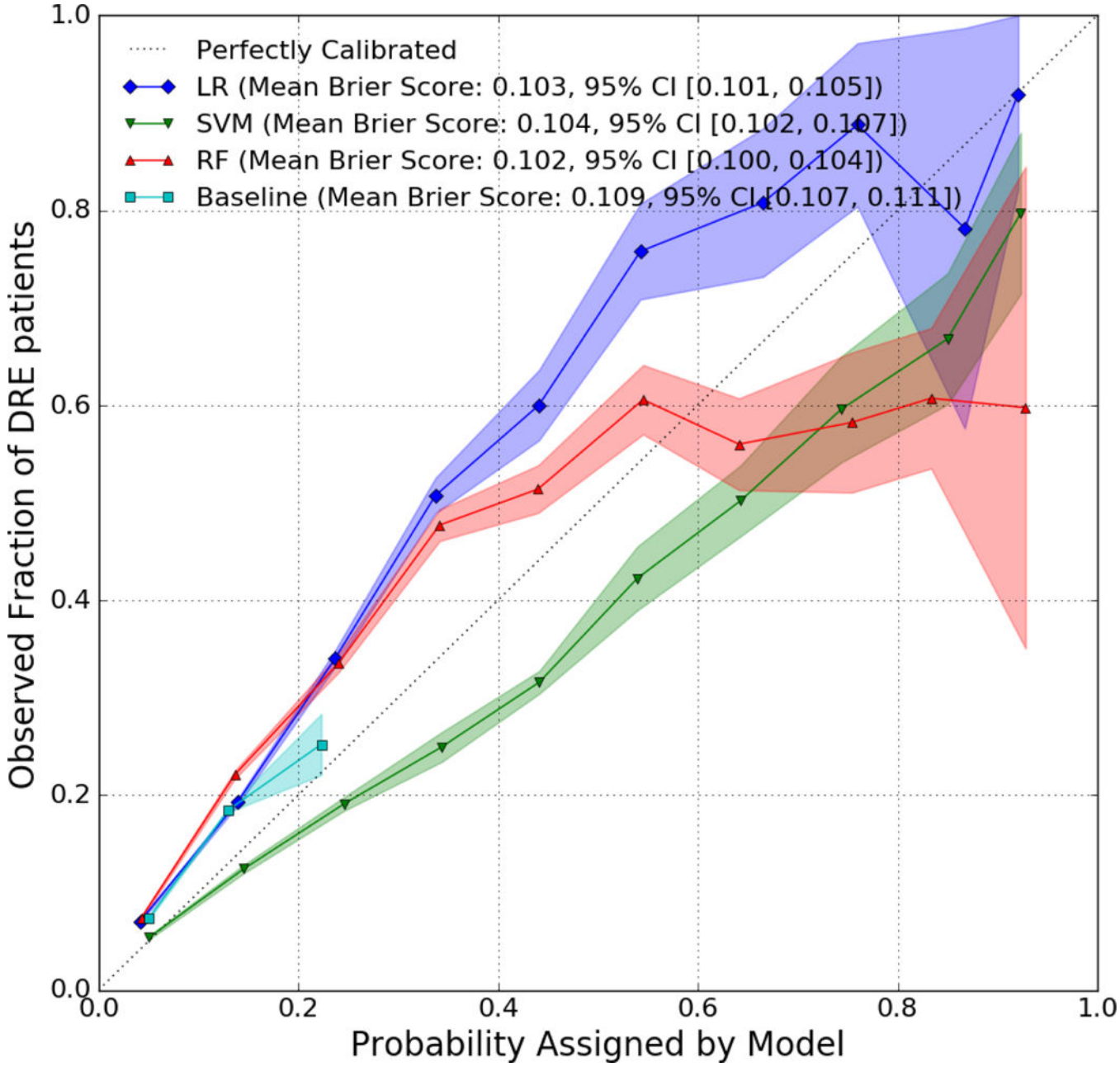
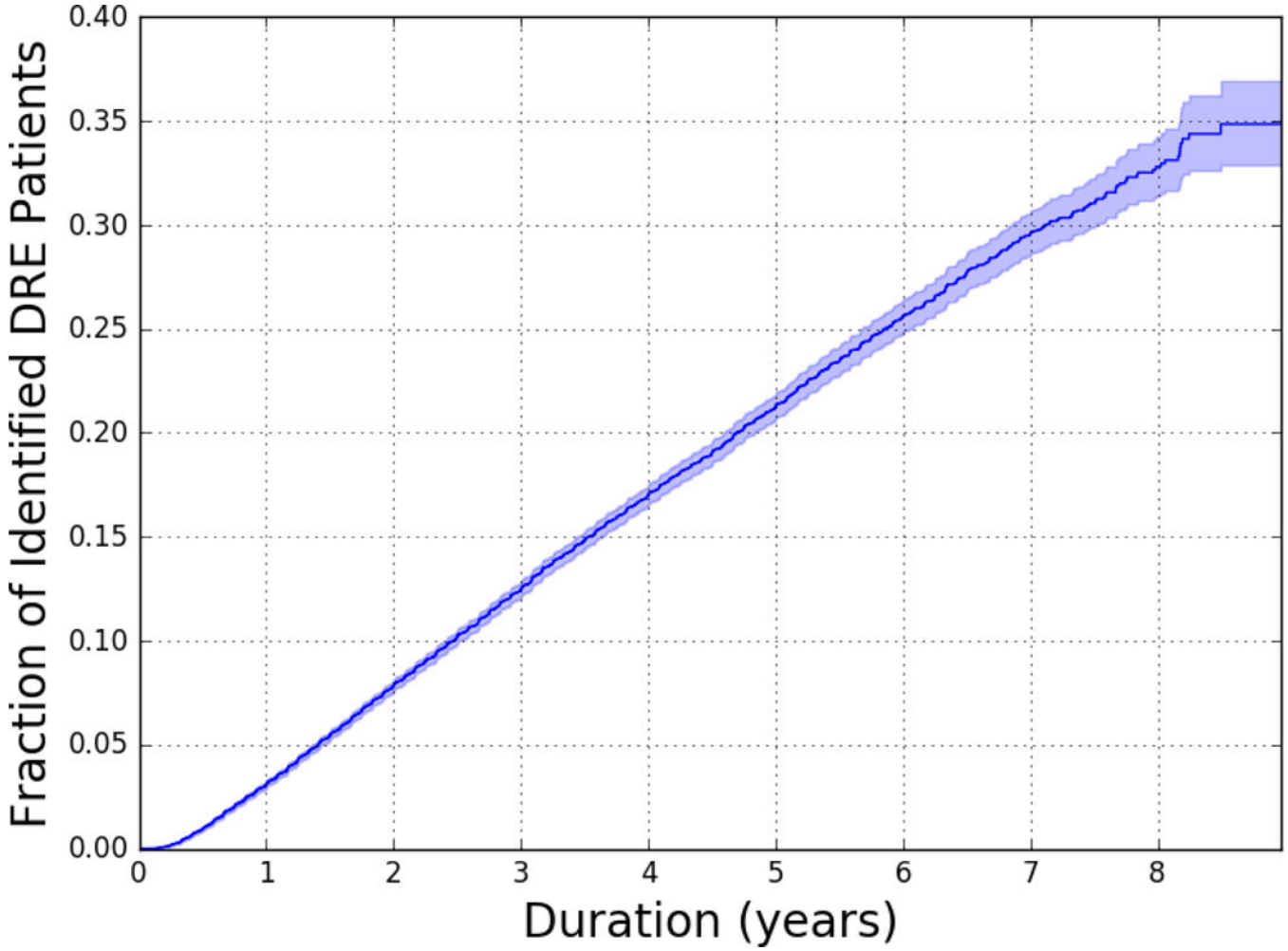


Figure 2. Risk Calibration curves for each model

Each curve is generated after the probability calibration step with isotonic regression. Brier score which lies between 0 and 1 for each model is also reported. A curve closer to the dotted diagonal line and has lower Brier score is the better calibrated model.



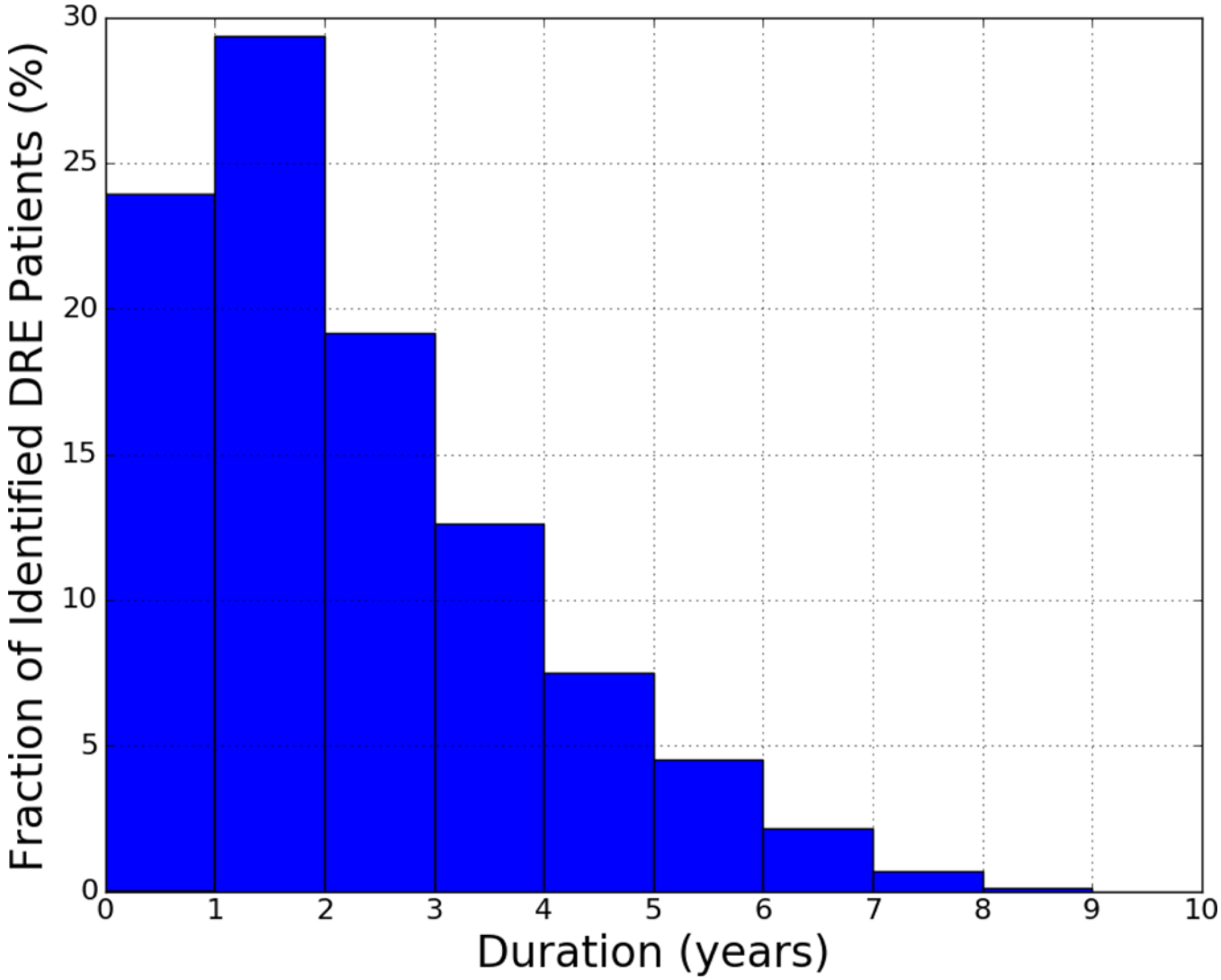


Figure 3. Nelson-Aalen cumulative hazard function curve for status quo.

x-axis represents the time taken to be classified as DRE patient after patients were prescribed the first AED and their corresponding y value depicts the fraction of DRE patients classified until that time by the definition of status quo.

Table 1.

Cohort characteristics

Metric	Count
No of Patients with Epilepsy	1,376,756
Age at Index Date	Mean: 44.6 Median: 46.4 IQR: 34.3
Gender	Male: 45.5% Female: 54.5%
No of Pharmacy Claims	28,403,939
No of Diagnosis Claims	173,570,273
No of Inpatient Encounters	201,202
No of Outpatient Encounters	707,332
No of ER Encounters	432,915
No of AEDs	20

Author Manuscript

Author Manuscript

Author Manuscript

Table 2 :

Impact of Inclusion & Exclusion criteria on the cohort size

Inclusion criteria	Percentage of patients retained
Epilepsy Diagnosis Criteria	100 %
Age Requirement	86.4 %
Data Completeness	63.7 %
Drug resistance – Non-drug resistant	38.7 %
Drug resistance – Drug resistant	3.6 %

Table 3:

Dataset Statistics

Type of Dataset / Class	Case	Control
Training	43,097	485,543
Hold out Test	6,819	46,799

Table 4:

Case-Control Statistics

	Case	Control
Age	Mean: 41.25 Median: 40.95 IQR: 23.43	Mean: 51.43 Median: 52.12 IQR: 29.51
Gender	Male: 42.25% Female: 57.75%	Male: 45.09% Female: 54.91%
Seizure Proxy (last 1 year)	14.04%	18.26%
Hypertension	15.20%	37.95%
Cardiac Arrhythmia	4.48%	12.36%
Brain Tumor	1.96%	3.43%
Solid Tumor	1.03%	4.20%
Anoxic Brain Injury	0.39%	1.40%
Myocardial Infarction	0.37%	1.59%

Table 5.

Top 20 predictive features.

Top 20 most predictive features among XX features in the best performing model

Feature (technical name given in parentheses)	Value type	MIN	MAX	MEDIAN	MEAN	STDEV
Indicators of level of activity within the medical system						
Number of claims prior to the past year with diagnoses classified as ““CCS 259” (“residual codes; unclassified”)	Count	0	1679	0	1.7	9.7
Number of months in the past year with claims containing any diagnostic code	Count	0	12	6	5.9	4.4
Number claims prior to the past year for diagnostic tests classified as “CCS 227” (“Other diagnostic procedures: interview; evaluation; consultation”)	Count	0	528	2	5.8	11.2
Number of claims in the past year for procedures classified as “CCS 227” (“Other diagnostic procedures: interview; evaluation; consultation)	Count	0	1281	7	17.1	28.3
Number of claims prior to the past year for radiology procedures classified as ““CCS 226” (“Other diagnostic radiology and related techniques”)	Count	0	135	0	1.6	3.7
Number of claims prior to the past year for chemistry and hematology tests (CCS 233: “Chemistry and hematology”)	Count	0	1535	0	5.6	20.0
Number of claims in the past month for any procedure	Count	0	643	0	1.9	5.7
Number of claims in the past year for any procedure	Count	0	2316	7	18.5	37.2
Number of claims prior to the past year for therapeutic procedures classified as “CCS 231” (“Other therapeutic procedures”)	Count	0	1140	0	2.3	8.6
Number months in the past year with claims for any prescription	Count	0	12	12	10.4	3.1
Indicators of Comorbidities						
Age	Ordinal	16.0	89.05	50.98	50.55	18.74
Charlson Comorbidity Index	Score	0	25	0	1.8	2.8
Number of claims prior to the past year for diagnosis of essential hypertension (CCS 98, “Unspecified essential hypertension”)	Count	0	1208	0	2.5	8.8
Number of claims at any past time for lipid-lowering drugs (USP Class: “Dyslipidemics HMG CoA Reductase Inhibitors”)	Count	0	14	0	0.5	1.0
Number of prescriptions at any time in the past for SSRIs or SNRIs	Count	0	18	0	0.8	1.4
Number of claims at any past time for GABA agonists (USP Drug Class: “Gamma-aminobutyric Acid (GABA) Augmenting Agents”)	Count	0	15	0	0.5	1.0
Indicators of Epilepsy and Epilepsy Complexity						
Number claims prior to the past year with an epilepsy diagnostic code (CCS 83: “Epilepsy; convulsions”)	Count	0	1594	1	3.4	10.2
Number of claims in the past year with an epilepsy diagnostic code (CCS 83: “Epilepsy; convulsions)	Count	0	368	0	1.5	4.6
Epilepsy Comorbidity Score	Score	0	24	0	1.5	2.4
First AED is prescribed by a neurologist.	Yes/no (1/0)	0	1	0	0.2	0.4

Legend: CCS, Clinical Classification Software (see: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/CCSUsersGuide.pdf); USP drug class, The U.S. Pharmacopeial Convention drug classification, (see: http://www.usp.org/sites/default/files/usp_pdf/EN/healthcareProfessionals/uspmmg_v6_0_cat-class.pdf); serotonin serotonin reuptake inhibitor; SNRI, serotonin-norepinephrine reuptake inhibitor; GABA, Gamma-aminobutyric Acid; HMG CoA reductase, 3-hydroxy-3-methyl-glutaryl-coenzyme A reductase; AED, anti-epileptic drug.