Using Machine Learning to Predict Schizophrenia Admittance

Pranav Warman[1], Gopalkumar Rakesh MD[2], Linda Adams[1], Beepul Bharti[1], Katherine Heller PhD[3], Jane Gagliardi MD, MHS, FACP, FAPA[2]

1-Trinity College of Arts and Sciences, 2- Department of Psychiatry and Behavioral Sciences, 3- Department of Statistical Sciences

Introduction

Abstract. Schizophrenia is a mental illness that affects ~1.1% of the US population and often requires lifelong treatment. Between 2009-2011, over 382,000 schizophrenic patients came in to the emergency department, and nearly 50% of them were later admitted as inpatients. Currently, due to the complexity of schizophrenia and the lack of a definitive biomarker, it is nearly impossible for a clinical provider to know prospectively which patients would benefit from more intensive resources including community support or clozapine. Here, we describe several models that are able to predict a patient's admittance as an inpatient from several simple and widely available data points taken during a patient's ER visit. Additionally, we note several interesting observations with regards to Duke's EHR and medication paths for schizophrenics.

Background.

Overall accuracy varies from 70 to 90 % in verifying the diagnosis of schizophrenia versus healthy controls based on applying machine learning (ML) to brain scans. Imaging datasets that machine learning has been applied to include structural MRI (measuring grey matter volumes), functional MRI using cognitive tasks measuring domains like working memory, white matter integrity (measured using fractional anisotropy) and electroencephalography (EEG) activity. There have also been attempts at correlating symptom dimensions with brain activity measured using electroencephalograph (EEG) or magnetoencephalography (MEG). The most common ML technique that has been used is support vector machine (SVM). To our knowledge there has been only one study looking at prediction of clinical course and outcome using administrative claims data. We lack clinical prediction tools that can stratify risk of relapse, and predict clinical course and prognosis for the disease.

Data Analysis

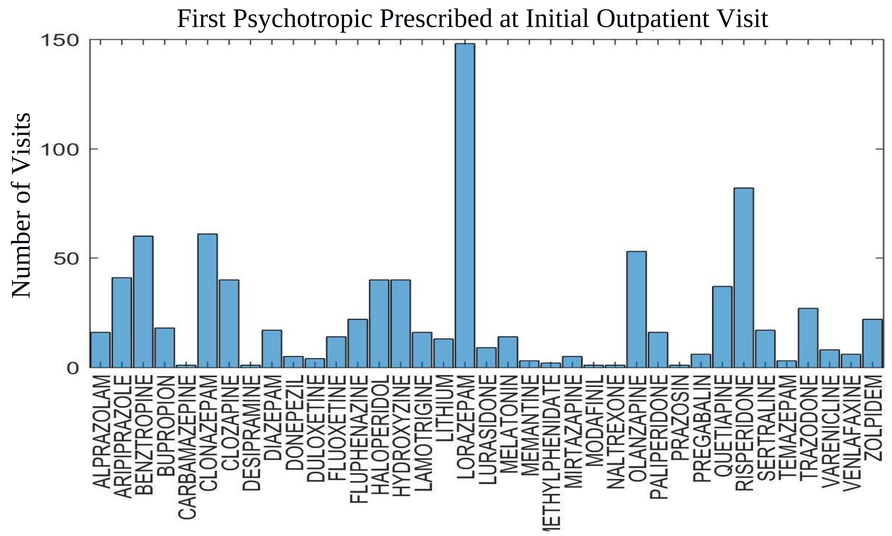
De-identified data was extracted from EPIC Maestro by an honest broker and provided to the team. Criteria for extraction was a diagnosis of schizophrenia and having had an inpatient, outpatient or ER visit over the course of time period 2014-2017 (1/1/2014- 4/10/17). Extracted fields included age, gender, diagnoses entered at each visit, medications at each visit, problem lists and labs.

Quick Overview:

- A total of 1351 patients were found to have inpatient, outpatient and emergency room (ER) visits.
- 1277 patients had only outpatient visits,.
- Of the 1277 patients, 330 outpatients had an emergency room or inpatient visit.
- A total of 1962 emergency room /inpatient total visits.
- In the inpatient dataset – M:F ratio is 729:622, outpatient dataset this is 682:594.
- Median age in the 2 populations are 49 and 51 years, respectively

Year	ER Visits/month	Unique visits/month
2014	2	2
2015	4	4
2016	4	3
2017	3	2

Data Analysis (Continued)



As we screened through the data we observed a wide heterogeneity in medications prescribed, as observed here with the array of first medications provided. Furthermore, **no two schizophrenic patients were prescribed the same order of medications.**

Additionally, the temporal aspect of modeling became dubious when it was noted that a patient's E**ncounter IDs could span years, making them non-sensical** and it impossible to distinguish patient visits from each other.

Machine Learning

Methods

Feature Matrix Construction:

1. Data sheet was reduced to only incorporate patients who had two or more visits, in order to increase density of information per patient.
2. Several different features were mined from the data sheet (e.g., medications prescribed, abnormal lab panels, insurance groups, etc.) and aggregated per patient
3. Due to the lack of a robust temporal dimension, the values per feature were not binned by month but rather totaled
4. Values were normalized between -3 and 3 where appropriate. (Note: features such as gender and insurance groups were defined values.)
5. Binary output values (i.e, whether the patient was admitted as an inpatient or in the ER) were mined and linked to the patient.

Machine Learning

1. A L1-Regularized Logistic Regression, Naive Bayes, and a Support Vector Machine with a Gaussian kernel were trained and tested on the feature matrix with a 70/30 split.

$$\hat{y} = \underset{k \in 1,2}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{9} p(x_i \mid C_k)$$

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

(a) Optimization of L1-Regularized Logistic Regression

$$p(y = 1|\mathbf{x};\theta) = \sigma(\theta^\top \mathbf{x}) = \frac{1}{1+\exp(-\theta^\top \mathbf{x})}$$

$$\min_{\theta} \sum_{i=1}^{754} -\log p(y^{(i)}|\mathbf{x}^{(i)};\theta) + \beta\|\theta\|$$

(b) Optimization of Naive Bayes Classifier

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$$

(c) Gaussian Kernel for the SVM

Results

Model	Train Accuracy (%)	Test Accuracy (%)
L1-Regularized Logistic Regression	72.22	76.59
Naive Bayes	67.94	74.04
Support Vector Machine + Gaussian Kernel	100	74.04

Accuracy results from running the predictive models on a 70/30 split of approximately 753 patients.

Discussion

Despite not yet being complete, the produced models have an accuracy above the 70th percentile in predicting whether a patient seen in the outpatient setting would need an emergency room visit or inpatient hospitalization. While this is a strong step forward, more work is needed in order to push this into clinics. Additionally, it appears that the SVM constructed needs to be re-paramterized as it's train accuracy signals overfitting. Once completed, though, these models hold promise for helping identify patients at highest risk for severe illness and therefore help triage them to more intensive outpatient services to decrease inpatient and ED use.

Future Work

- Explore the already built models to determine and understand which features are highly indicative for admittance
- Leverage tools such as TensorFlow to continue analyzing the dataset with models such as a RNN.
- Explore additional pre-processing techniques and time-series analysis.
- A dashboard software interface that calculates risk prediction based on inputs and stratifies patients from the first visit would be the endpoint.

Acknowledgments

The presenters gratefully acknowledge funding made available for this research by the Bass Family to the Duke University Social Science Research Institute and would like to thank Joe Futoma PhD, Mark Sendak MD, and Anmol for their guidance.

Sources

Zarogianni E, Moorhead TW, Lawrie SM. Machine learning approaches: from theory to application in schizophrenia. Neuroimage Clin. 2013 Sep 13;3:279-89.
Veronese E, Castellani U, Peruzzo D, Bellani M, Brambilla P. Machine learning approaches: from theory to application in schizophrenia. Comput Math Methods Med. 2013.
Wang Y, Iyengar V, Hu J, Kho D, Falconer E, Docherty JP, Yuen GY. Predicting Future High-Cost Schizophrenia Patients Using High-Dimensional Administrative Data. Front Psychiatry. 2017 Jun 30; 8:114.