Predictive Analytics Case Study of:

"A machine learning framework to predict the risk of opioid use disorder"

Author: Alysen Casaccio

April 14, 2024

Introduction

Problem Statement:

The authors of my selected article, "A Machine Learning Framework to Predict the Risk of opioid use disorder," felt that too little was understood about how large-scale data analytics could be utilized to help healthcare providers predict whether a given patient would develop opioid use disorder. The authors proposed a machine learning framework for identifying potential risk factors of opioid use disorder from large-scale healthcare claims data.

Problem Importance:

The opioid crisis has become a significant public health concern in the United States (Md Mahmudhul Hasan, 2021). Overdose deaths linked to prescription opioids take an average of 128 lives in the US every day (CDC, 2020), and opioid overdose deaths were four times higher in 2018 than in 1999 (CDC, 2020).

Opioids continue to be used as a common pain management intervention for many patients, and it is challenging for providers to assess patient potential for opioid dependence. As many as 25% of patients receiving long-term opioid therapy will struggle with opioid use disorder (CDC, 2020).

Data Acquisition:

The authors leveraged the Massachusetts All Payor Claim Datasets (MA APCD), housed at the Northeastern Center for Health Policy and Healthcare Research and overseen by the Center for Health Information and Analysis (CHIA). This database includes all medical claims, pharmacy claims, and member eligibility information associated with commercial insurance claims in Massachusetts between 2011 and 2015 (Md Mahmudhul Hasan, 2021). Due to a 2017 policy change issued by the U.S. Department of Health and Human Services (HHS), CHIA was required to remove all medical claims related to drug dependence after 2013. As a result, the authors focused their study cohort only to include patients from 2011 to 2013.

The dataset provided unique identifiers for all patients and settings of care or types of providers. These were used to link claims and individuals across files. The pharmacy claim file contains patient demographics, prescriber characteristics, fill dates, medication type, quantity, and days supply (Md Mahmudhul Hasan, 2021). The medical claim file included information about each patient's clinical history, including principal diagnoses, services, procedures, and payment (Md Mahmudhul Hasan, 2021). Both files could then be linked to the member eligibility file that contains information on the patient's age, gender, and insurance status (Md Mahmudhul Hasan, 2021).

Methods and Results

Data Preparation:

The authors prepared the data and made it suitable for training and testing models by aggregating claim-level information into patient-level information. Specific adjustments for missing values had to be made, including missing age values among the patients without opioid use disorder. These were replaced by the mean age of all other patients without opioid use disorder. Age was also treated as a categorical feature rather than continuous and five different bins were generated (18-25, 26-35, 36-55, 56-64, and above 65). This was divided after considering the overall age distribution (Md Mahmudhul Hasan, 2021). This was largely in alignment with another study (Brat, 2018), with adjustment for the low number of patients in the Massachusetts database under the age of 18 (these were excluded).

Records with missing gender were removed from the dataset, categorical ICD-9 codes were replaced with the description of each code to make the data and results easier to interpret, and an additional feature was engineered titled "Degree of Chronicity of Opioid Usage" (Md Mahmudhul Hasan, 2021). This was a categorical feature where the proportion of days covered (PDC) was used to determine the severity of the chronic nature of the usage (Md Mahmudhul Hasan, 2021). Because PDC was defined as the fraction of days a patient was on opioid medications within one year since their first opioid prescription, the authors were able to categorize this value into four distinct levels. Patients with less than

20% PDC were considered “non-chronic,” whereas those with a PCD value greater than 80% were considered “highly chronic,” two additional bins in the moderate range were also created.

The final analytic file contained over 600,000 patients and 12,000 features. Due to the high dimensionality of their dataset, they used the variance threshold (VT) method to remove features with low variance. Only features with a variance higher than the pre-defined threshold of 0.03 were retained. After the VT methodology, the number of features was reduced to 3,076 and the authors were confident that all had a variance of 0.03 or greater. Once this was complete, a Chi-squared test was implemented to identify and retain only the statistically significant features (>0.05 p-value). This further reduced the number of features to 2,628; the dataset was used to train the models (Md Mahmudhul Hasan, 2021).

Problem Solution:

The authors state that although clinicians may over-prescribe opioids for pain management, there is a legitimate need for prescribing opioids to individuals for certain clinical situations. A significant number of patients could have risk factors for opioid use disorder but still require opioids for pain management. This requires stronger tools to be available for providers to weigh risks and benefits. This study could represent a step toward developing these kinds of tools (Md Mahmudhul Hasan, 2021).

Modeling Techniques Used:

The authors used multiple modeling techniques, including recursive feature elimination (RFE) and the synthetic minority oversampling technique (SMOTE), and four different model types, including logistic regression, decision tree, random forest, and gradient boosting. All these techniques and the reasons they chose them are explained below.

Methodology Explained:

The first methodology used was recursive feature elimination (RFE). This methodology was needed due to the high dimensionality of the authors’ dataset. This technique eliminates features recursively by pruning the original set of features from the model(Md Mahmudhul Hasan, 2021). The

RFE takes a model trained with all available features and assigns weights as coefficients or feature importance. The least important features are pruned in each iteration based on weights. Using a cross-validation technique, the authors attempted to remove 10% of the features after each iteration (Md Mahmudhul Hasan, 2021).

Four different models (logistic regression, decision tree, random forest, and gradient boosting) were recursively trained in this way. To set a stopping condition for the RFE, they investigated the features that could achieve a higher area under the receiving operating characteristics curve (AUC) and determined the maximum AUC values for each predictive model.

Their final modeling technique was to handle the class imbalance in the dataset. In the study sample, only 1% of the patients were identified with opioid use disorder, which was a highly class-imbalanced sample. They used a Synthetic Minority Oversampling Technique (SMOTE) to over-sample the minority class (opioid use disorder patients). This technique was used to generate "synthetic" examples of a minority class, which resulted in a similar distribution of opioid use disorder patients and non-opioid use disorder patients in the dataset (Md Mahmudhul Hasan, 2021).

Metrics and Evaluation:

The metrics used to evaluate the model performance at each feature elimination step were precision, recall, F1-score, and AUC value. Recall and AUC, specifically, were selected to choose the final model based on two reasons: one, the authors wanted to identify as many patients with opioid use disorder as possible due to the patient risk of going without appropriate interventions if misclassified, and two, the AUC is an effective indicator of the model's ability to distinguish a rare class from the prevalent one (Md Mahmudhul Hasan, 2021).

Conclusion

Model Implementation:

The authors presented a comparative analysis across all four models. Based on the AUC values as the primary evaluation metric, the results were highest for Random Forest and Gradient Boosting, with

slightly lower rates on the Decision Tree model. All tree-based models outperformed the Logistic Regression model (Md Mahmudhul Hasan, 2021).

This study created a new machine learning-based framework using the healthcare administrative claims dataset to discover demographic and clinical features that predict opioid use disorder (Md Mahmudhul Hasan, 2021).

Actionable Consequences:

One common concern in using administrative claims is the lack of adequate clinical details for developing efficient and reliable predictive models for patient outcomes. However, the authors state that their study clearly demonstrated that this approach is feasible and clinically significant (Md Mahmudhul Hasan, 2021).

Lessons Learned:

Although the authors expressed many areas in which they attempted to overcome the limitations of prior studies, they also identified a series of limitations of their own study. Due to the lack of clinical history before 2011 and after 2013, they could not investigate how many patients were diagnosed with opioid use disorder in the future. In addition, they mention that the dataset is focused on only those covered under commercial insurance.

Future Approaches:

Additional future studies could expand on this work by implementing the author's framework on claims datasets in other states trying to get improvement on the opioid crisis. In addition, expanding the dataset into more vulnerable populations, like Medicare or Medicaid beneficiaries provides an additional approach.

The authors seek to develop this framework into a user interface so providers can use it as a ready-to-use tool in practice that would generate a risk score for an individual patient before taking a prescribing action at the point of care. They anticipate that such a decision support tool would influence

the current standard of practice and promote informed access to opioids for patients at risk (Md Mahmudhul Hasan, 2021).

References

Brat, G. A. (2018). Postsurgical prescriptions for opioid naive patients and association with overdose and misuse: a retrospective cohort study. *BMJ*, 360.

CDC, C. f. (2020, May 11). *Opioid overdoes, drug overdose death data and understanding the epidemic*. Retrieved from CDC.gov: www.cdc.gov/druoverdose/data/statedeaths.html

Kyle Gallatin, C. A. (2023). *Machine Learning with Python.* Sebastopol, CA: O'Reilly Media, Inc.

Md Mahmudhul Hasan, G. J.-E.-A. (2021). A machine learning framework to predict the risk of opioid use disorder. *Machine Learning with Applications*, 1-15.