**Introduction**

Start by describing the imbalanced learning problem, similar to that of Maloof 2003, but talk about it in the context of clinical/public health problems (for example, talk about postoperative complications – Katie will do this – and chronic opioid therapy, and perhaps others, like sepsis, and maybe examples in genomics).

Learning with imbalanced data has always led to a classification issue. The predictive performance of models deals with issues of low sensitivity when the outcome is rare. Imbalanced data can be found anywhere, and is often encountered in the clinical setting when diseases and diagnoses are only found in a small portion of the population. For example, sepsis incidence has been shown to be 3 cases per 1000 population or 2.26 cases per 100 hospital discharges (1). When attempting to predict an outcome this rare, the sensitivity—or proportion of those with the outcome correctly identified—is critical, yet the imbalanced nature of the data makes this difficult as algorithms tend to learn how to predict the majority class much better than the minority.

\*\*do I discuss specifically chronic opioid therapy or just opioids in general, and can I say here or do I wait to say that we are using a COT dataset? Otherwise I’m not sure which numbers to report in terms of prevalence or something, because stats are so specific to one thing regarding opioids\*\*

Next, discuss statistical methods people typically use to handle imbalanced data sets: sampling, ROC, Youden’s J, etc.

Many people have tried to combat this issue in the past. Other researches have attempted to fix the issue of over fitting with shrinkage methods like ridge regression or lasso [^fn8], random forests [^fn9], as well as many sampling techniques [^fn10] and adjusting probability cut-points with ROC analyses. [^fn11]

\*\*Do I need more detail here?

Brief paragraph of what we will show in this paper.

However, no one has yet explored the trade offs of probability cut-points and sampling methods, which is what we desired to do in this analysis to achieve the best prediction. We sought to identify rare outcomes in two case studies using several different approaches to optimize probability cut-points and sample the data in order to compare and contrast the predictive performance of these imbalanced data techniques.

\*\*Can we say no one yet explored the trade offs because of the other similar paper?

**Methods**

Data: We will show the tradeoffs for different methods in two separate real clinical data sets and one simulation study.

\*\*Sums it up so concisely don’t know what else to add here

Statistical methods: describe the methods we will use. I believe for all we use lasso and a combination of sampling and probability thresholds.

The model used for this analysis was cross-validated lasso regression ran in R. [^fn3][^fn4] This was chosen as it has been found to perform better predictor selection than stepwise selection, and as we were not interested in having interpretable coefficients. We used a roughly 2/3rd temporal split of the data to create training and testing datasets

As a comparison, we first evaluated the prediction performance of the dataset without sampling to see the effects of the imbalanced data on the accuracy, sensitivity, and specificity. This was to serve as a baseline to compare with the techniques available to mitigate the issue of poor sensitivity. The predicted probability cut-point used here was rounding at the standard 0.5 that would be appropriate in balanced datasets.

The first approach used to improve performance was choosing a more informed probability cut-point for the data. This was done using the Youden Index [^fn5], which finds the maximum of the receiver operating characteristic (ROC) curve [^fn6] with the pROC package. The second approach was through sampling the dataset. Three types of sampling methods were compared--down sampling, up sampling, and Synthetic Minority Over-sampling Technique (SMOTE)[^fn7]. Down sampling takes a random sample from the majority class, in this case those who are not classified as having chronic opioid therapy, in order to match the size of the minority class. Up sampling does the reverse to take random samples of the minority class in order to match the majority. SMOTE combines sampling both from the majority and minority, but instead of taking identical copies of the minority it creates synthetic observations using nearest neighbors. For each of the three sampling techniques, the standard 0.5 probability cut-point was used and the cut-point was optimized using the Youden Index for comparison, yielding a total of six models.

**Case Study 1: Risk factors for chronic opioid therapy**

Write a paragraph describing the opioid study then present results.

To illustrate this issue with our first case study, we are using electronic health record data from Denver Health from the years 2008 to 2014 of patients for patients with chronic opioid therapy (COT). This is an urban, safety-net hospital. Five percent of the 27,705 patients were reported with the outcome, which was defined as receipt of $\geq$ 90-day supply of opioids with < 30-day gap in supply over a 180-day period or receipt of $\geq$ 10 opioid prescriptions over one year. The data also contained demographic information on the patient including age, race, gender, history of chronic pain, and length of hospital stay. Years 2008-2011 were used to train (65%), and 2012-2014 were used to test (35%). The predictors were narrowed from 50 to 35 based on clinical relevance.

As expected, without using an optimized cut-point or sampling technique, the sensitivity of the model was extremely poor at 8%, with high specificity and accuracy (99% and 96%). Simply choosing a more informed probability cut-point at 0.043 instead of 0.5 improved the sensitivity to 85% and brought the specificity down to 73%. This cut-point is intuitive as the outcome is present at 5% in the dataset, which would be consistent with a 0.5 cutoff in an evenly split dataset. The up and down sampled datasets both showed the same improved sensitivity with Youden index for probability cut-points at about 0.4, also with close specificities of 74 and 73%. SMOTE on the other hand, resulted in 74% sensitivity and 84% specificity. However, there were improvements in accuracy for SMOTE at 86% as compared to the other three approaches, which had accuracies at 86-87%. When using the standard 0.5 cut-point for up and down sampling, the accuracy improved from 74% to 82 and 81%, while the values for sensitivity and specificity stayed high but reversed in magnitude with specificities in the 70 percent range and sensitivities in the 80s. There was no change to the negative predicted value across the approaches, and a decrease in positive predicted value. In terms of the ROC analysis, the area under of the curve for each approach was about the same at 86-87%. See Table 1 for full results for the cut-point, sensitivity, specificity, accuracy, negative predicted value, positive predicted value, and area under the curve.

**Case Study 2: Identification of surgical site infections**

Katie: write a paragraph describing the surgical infections study then present results.

**Simulation study**

Describe our simulation study and present results.

In order to explore how these techniques performed at a larger range of prevalences, we ran a simulation study for data with 1, 3, 5, 10, 25, and 50 percent outcome. We used the chronic opioid therapy dataset by selecting three of some of the most significant predictors—age, chronic pain at discharge, and receipt of opioid at discharge. We first ran a logistic regression with these three variables in order to choose our coefficients for the simulated data. Then we simulated using a logistic distribution to create the new datasets, changing the prevalence by adjusting the intercept. From there, we employed the same methods as in the case studies to get results for sensitivity, specificity, and accuracy with various sampling and cutpoints.

\*\*How much detail to go into for describing how we created the simulation?

The results of this simulation study proved the linear one to one relationship of the Youden Index versus prevalence. At every prevalence, the ideal cut-point based off of the Youden Index was equal to the prevalence—e.g. 5% outcome, 0.05 cut-point.

\*\*Don’t know if we’re sure about sensitivity versus prevalence results right now

\*\*I couldn’t possibly go into the details of all the sensitivity and specificities of each model for each prevalence; do I mention any of it or just say refer to table?

**Discussion**

Discuss the findings from each and how they compare or contrast and what we learned from the simulation study. Conclude with some remarks about how we’ve shown a process for handling imbalance and that context matters, so there is no recipe for the best approach – one might want to follow a process similar to what we did.

\*\*Holding off on the discussion right now until we have all the final results from surgical and simulation

Added publication reference in intro

1. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. https://www.ncbi.nlm.nih.gov/pubmed/11445675/