Thesis Proposal

Predictive Modeling with Imbalanced Data

Alyssa Forber
Susan Calcaterra, Elizabeth Juarez-Colunga, Kathryn Colborn

December 06, 2017

Introduction

The misuse of opioids has become an increasing problem in the United States, and the need to identify patients at increase risk for opioid abuse is immediate. The number of opioid related overdose deaths has continued to increase, and deaths from prescripted opioids have more than quadrupled since 1997.[1] In 2015 alone there were 20,101 overdose deaths related to prescription opioids and 12,990 overdose deaths related to heroin.[2]

This issue now presents a statical problem, as this is a rare outcome which makes prediction more challenging. Predictive modeling with imbalanced data has been found to report low sensitivity (reference?). Many people have tried to combat this issue in the past. GIVE EXAMPLES.

- cite some other issues in health with rare outcomes and what people have done
- give fair bit of background (stats in medical, etc) find where people try to predict rare outcomes

However, no one has yet explored the tradeoffs of probility cut-points and sampling methods, which is what we desired to do in this analysis to achieve the best prediction. We sought to identify patients at risk for chronic opioid therapy 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.

Methods

Data

To illustrate this issue, 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

[1]CDC. Wide-ranging online data for epidemiologic research (WONDER). Atlanta, GA: CDC, National Center for Health Statistics; 2016. Available at http://wonder.cdc.gov.

[2]Rudd RA, Seth P, David F, Scholl L. Increases in Drug and Opioid-Involved Overdose Deaths — United States, 2010–2015. MMWR Morb Mortal Wkly Rep 2016;65:1445–1452. DOI: http://dx.doi.org/10.15585/mmwr.mm655051e1

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.

Statistical Methods

The analysis was done in RStudio version 1.1.383. We used a roughly 2/3rd temporal split of the data to create training and testing datasets, where years 2008-2011 were used to train (65%), and 2012-2014 were used to test (35%).

The model used for this analysis was cross validated lasso regression.[3][4] This was chosen as it has been found to perform better predictor selection than stepwise selection (reference on this?), and as we were not interested in having interpretable coefficients.

The predictors were first narrowed from 50 to 35 based on clinical relevance.

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[5], which finds the maximum of the receiver operating characteristic (ROC) curve[6] 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)[7]. 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 (reference?). Up sampling does the reverse to take random samples of the minority class in order to match the majority (reference?). 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 (reference). 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.

[3]Browne, M. W. (2000). "Cross-validation methods." Journal of Mathematical Psychology 44(1): 108-132.

[4]Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. J. Roy. Stat. Soc. Ser. B, 58(1), 267–288.

[5]Fluss, R., et al. (2005). "Estimation of the Youden Index and its associated cutoff point." Biom J 47(4): 458-472.

[6]Zou, K. H., et al. (2007). "Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models." Circulation 115(5): 654-657.

[7]Nakamura, M., et al. (2013). "LVQ-SMOTE - Learning Vector Quantization based Synthetic Minority Over-sampling Technique for biomedical data." BioData Min 6(1): 16.

Results

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 to 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 a evenly split dataset. The up and down sampled datasets both showed the same improved sensitivity with 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%.

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 2 for full results for the cut-point, sensitivity, specificity, accuracy, negative predicted value, positive predicted value, and area under the curve.

Conclusions

Both sampling and optimizing probability cutpoints, alone and combined, improved the predicted sensitivity drastically. There was not a distinguishable difference between over and under sampling for either approach, while the SMOTE algorithm appeared to give lower sensitivity than the other two sampling techniques. In addition, using the Youden index resulted in very similiar results as to when the Youden index was used in conjuction with sampling. This may suggest that using both sampling and an informed cut-point at the same time is not necessary, since equivalant predictive performance is seen with the Youden index alone.

Moreover, the decision as to what approach to use is largely influenced by the circumstances of the problem. In some cases it may be more important to have a higher sensitivity than specificy, or vice versa. It may come down to the clinician's decision to have a threshold of sensitivity that is needed. As it can be a case by case choice with many tradeoffs, no one method is recommended above the others.

To expland upon this analysis, we plan to implement a simulation that explores predictive performance of these models at different rarities for the outcome. This is to get a better idea of how they perform in other cirmunstances as well as see when these techniques are no longer needed because the imbalance does not affect model performance. We also plan to implement bootstrap aggregating, also known as bagging. This model averaging approach will hopefully improve the stability and accuracy of our models. Lastly, other sampling or optimzed probability cut-point techniques that are found to be promising may be incorporated into the analysis.

Table 1: Results

Data	Threshold	Specificity	Sensitivity	NPV	PPV	Accuracy	AUC	Covariates
Unsampled 0.5	0.5	99	8	96	35	96	86	31
Unsampled	0.043	73	85	99	12	73	86	31
Down sampled 0.5	0.5	81	75	99	15	81	86	34
Down sampled	0.401	73	85	99	12	74	86	34
Up sampled 0.5	0.5	82	75	99	15	82	87	34
Up sampled	0.399	74	85	99	12	74	87	34
SMOTE 0.5	0.5	86	71	99	17	85	86	33
SMOTE	0.472	84	74	99	17	84	86	33

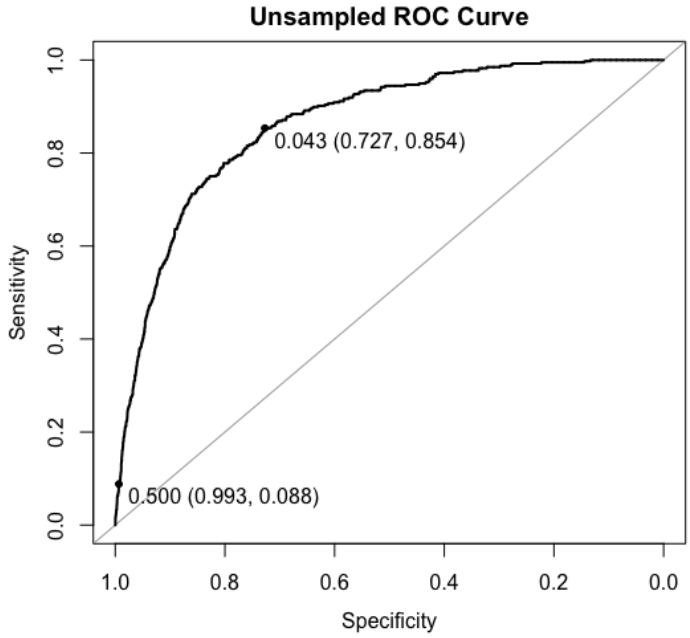

Figure 1: ROC for Unsampled Data: Younden and 0.5 cutoffs

References

Appendix

		Chronic Opioid Use		P-Value
	Total (N=27,705)	Yes (N=1,457)	No (N=26,248)	
Gender, n (%)				
Female	12,933 (46.7)	652 (44.7)	12,281 (46.8)	0.13
Race, n (%)				
Hispanic	10,798 (39.0)	580 (39.8)	10,218 (38.9)	0.01
Non-Hispanic White	10,645 (38.4)	555 (38.1)	10,090 (38.4)	
African American	4,842 (17.5)	273 (18.7)	4,569 (17.4)	
Other or Unknown	1,420 (5.1)	49 (3.4)	1,371 (5.2)	
Age at Index Admission (Years), n (%)				
15-<35	6,017 (21.7)	150 (10.3)	5,867 (22.4)	<0.0001
35-<45	4,734 (17.1)	267 (18.3)	4,467 (17.0)	
45-<55	6,919 (25.0)	506 (34.7)	6,413 (24.4)	
55-<65	5,880 (21.2)	400 (27.5)	5,480 (20.9)	
65-<75	2,745 (9.9)	110 (7.5)	2,635 (10)	
75-ł85	1,410 (5.1)	24 (1.6)	1,386 (5.3)	
Mean (SD)	48.1 (16.0)	50.2 (11.6)	48.0 (16.2)	
Median (25th, 75th)	49 (37, 59)	51 (43, 58)	49 (36, 59)	
Insurance Status, n (%)				
Discount Payment Plan*	8,499 (30.7)	576 (39.5)	7,923 (30.2)	<0.0001
Medicaid	8,575 (31.0)	531 (36.4)	8,044 (30.6)	
Medicare	6,260 (22.6)	259 (17.8)	6,001 (22.9)	

		Chronic Opioid Use		P-Value
Commercial	2,402 (8.7)	50 (3.4)	2,352 (9.0)	
Other/Unknown/Self-Pay	1,969 (7.1)	41 (2.8)	1,928 (7.3)	
Three Year History of, n (%)				
Tobacco Use Disorder	9,682 (34.9)	716 (49.1)	8,966 (34.2)	<0.0001
Alcohol Use Disorder	7,167 (25.9)	408 (28.0)	6,759 (25.8)	0.06
Stimulant Use Disorder	1,719 (6.2)	118 (8.1)	1,601 (6.1)	0.003
Opioid Use Disorder	672 (2.4)	44 (3.0)	628 (2.4)	0.13
Chronic Pain	14,914 (53.8)	1,105 (75.8)	13,809 (52.6)	<0.0001
Acute Pain	10,073 (36.4)	611 (41.9)	9,462 (36.0)	<0.0001
Top 3 Mental Health Disorders, n (%)				
Depression	6,318 (22.8)	491 (33.7)	5,827 (22.2)	<0.0001
Anxiety Disorder	3,677 (13.3)	265 (18.2)	3,412 (13.0)	<0.0001
Bipolar Disorder	2,362 (8.5)	135 (9.3)	2,227 (8.5)	0.3
Any Mental Health Disorder n (%)	9,805 (35.4)	634 (43.5)	9,171 (34.9)	<0.0001
Top 3 Chronic Medical Conditions, n (%)				
Hypertension	11,799 (42.6)	773 (53.1)	11,026 (42.0)	<0.0001
Respiratory Disease	7,060 (25.5)	444 (30.5)	6,616 (25.2)	<0.0001
Diabetes Mellitus	5,701 (20.6)	376 (25.8)	5,325 (20.3)	<0.0001
Any Chronic Medical Condition, n (%)	17,535 (63.3)	1,102 (75.6)	16,433 (62.6)	<0.0001
Charlson Comorbidity Index from 3 Year Diagnosis History				

		Chronic Opioid Use		P-Value
Mean (SD)	1.9 (2.2)	2.4 (2.5)	1.9 (2.2)	<0.0001
Median (25th, 75th)	1 (0, 3)	2.0 (1, 3)	1.0 (0, 3)	
Discharge Diagnoses, n (%)				
Chronic Pain_	8,346 (30.1)	729 (50.0)	7,617 (29.0)	<0.0001
Acute Painŕ	4,586 (16.6)	255 (17.5)	4,331 (16.5)	0.32
Neoplasm¤	1,447 (5.2)	170 (11.7)	1,277 (4.9)	<0.0001
Top 3 Surgical Procedures During Initial Hospitalization, n (%)				
Digestive System	3,437 (12.4)	225 (15.4)	3,212 (12.2)	<0.001
Musculoskeletal System	3,037 (11.0)	258 (17.7)	2,779 (10.6)	<0.0001
Cardiovascular System	2,312 (8.3)	157 (10.8)	2,155 (8.2)	<0.001
Patients Who Had Surgical Procedure During Index Hospitalization, n (%)	10,956 (39.5)	700 (48.0)	10,256 (39.1)	<0.0001
Number of Healthcare Encounters in the One Year Preceding the Index Admission, n (%)				
0	23,280 (84.0)	1,196 (82.1)	22,084 (84.1)	0.03
1	3,413 (12.3)	197 (13.5)	3,216 (12.3)	
2+	1,012 (3.7)	64 (4.4)	948 (3.6)	
Mean (SD)	0.2 (0.6)	0.2 (0.7)	0.2 (0.6)	
Median (25th, 75th)	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)	
Past Year Benzodiazepine Receipt, n (%)	1,606 (5.8)	227 (15.6)	1,379 (5.3)	<0.0001
Past Year Receipt of Non-Opioid Analgesics (NSAIDs, neuropathic agents, topical capsaicin & lidocaine), n (%)	4,875 (17.6)	620 (42.6)	4,255 (16.2)	<0.0001
Past Year Number of Opioid Prescriptions Filled, n (%)				

		Chronic Opioid Use		P-Value
0	21,543 (77.8)	549 (37.7)	20,994 (80.0)	<0.0001
1	3,167 (11.7)	249 (17.1)	2,918 (11.1)	
2	1,331 (4.8)	197 (13.5)	1,134 (4.3)	
3	646 (2.3)	132 (9.1)	514 (2.0)	
9-Apr	1,018 (3.7)	330 (22.6)	688 (2.6)	
Receipt of Opioid at Discharge, n (%)	8,028 (29.0)	817 (56.1)	7,211 (27.5)	<0.0001
Milligrams of Morphine Per Hospital Day, n (%)				
0	9,655 (34.8)	189 (13.0)	9,466 (36.1)	<0.0001
0.01 < 10	3,320 (12.0)	108 (7.4)	3,212 (12.2)	
10 < 51	7,337 (26.5)	490 (33.6)	6,847 (26.1)	
51 < 100	4,413 (15.)	371 (25.5)	4,042 (15.4)	
100+	2,980 (10.8%)	299 (20.5)	2,681 (10.2)	
Mean (SD)	37.7 (65.4)	64.4 (76.7)	36.2 (64.4)	
Median (25th, 75th)	12.5 (0, 54.7)	45.5 (14.3, 90.2)	10.8 (0, 52.2)	
Length of Hospital Stay (days)				
1	8,449 (30.0)	383 (26.3)	8,066 (30.7)	0.0003
2	5,655 (20.4)	282 (19.4)	5,373 (20.5)	
5-Mar	7,801 (28.2)	450 (30.9)	7,351 (28.0)	
6+	5,800 (20.9)	342 (23.5)	5,458 (20.8)	
Mean (SD)	4.7 (9.0)	4.9 (7.7)	4.6 (9.1)	
Median (25th, 75th)	2 (1, 5)	3 (1, 5)	2 (1, 5)	

		Chronic Opioid Use		P-Value
Number of Subsequent Hospitalizations within 12 Months post Hospital Discharge				<0.001
Mean (SD)	NA	1.48 (2.20)	0.54 (1.21)	
Median (25th, 75th)	NA	1 (0,2)	0 (0,1)	