Thesis Proposal

Predictive Modeling with Imbalanced Data

Alyssa Forber

University of Colorado, Anschutz Medical Campus

December 2017

Outline

Motivation

- Chronic opioid therapy has become an epidemic
- Over 2 million people had a prescription opioid use disorder (2015 National Survey of Drug Use and Health)
- Important to identity patients at high risk
- Allow for hospitals to make informative decisions about prescribing opioids

Imbalanced Learning Problem

- Presents a problem of imblananced data
- Poor sensitivity with rare outcomes
- Need to improve predictive performance

The Data

- Design: Denver Health retrospective analysis electronic health record (EHR) data from 2008 to 2014.
- Patients: Hospitalized patients at an urban, safety-net hospital.
- Definition of Chronic Opioid Therapy (COT) one year following the index hospital discharge:

  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.

Patient Population

- 27,705 patients
- Majority had incomes <185% of the Federal Poverty Level
- 70% were ethnic minorities
- 5% with COT
- Excluded Patients:
  - <15 or >85 years old
  - Those in prison, jail, or police custody
  - Those who died within one year following their index hospitalization
  - Patients with <2 healthcare visits to Denver Health three years preceding their index hospitalization
  - Undocumented persons receiving emergent hemodialysis
  - Obstetric patients

Table 1

Variable	**Yes COT**	**No COT**	**p-value**
	1,457 (5%)	26,248 (95%)	
Age 15-35	10%	22%	$<.001$
Age 45-55	35%	24%	$<.001$
Age 55-65	28%	21%	$<.001$
Discount payment or Medicaid	76%	61%	$<.001$
History of chronic pain	76%	53%	$<.001$
Discharge diagnosis chronic pain	50%	29%	$<.001$
Surgical patient	48%	39%	$<.001$
Past year:			
Benzodiazepine	16%	5%	$<.001$
Non-opioid analgesics	25%	9%	$<.001$
Number of opioid prescriptions:			
0	38%	80%	
1	17%	11%	
2	14%	4%	
3	9%	2%	
4-9	23%	3%	$<.001$
Receipt of opioid at discharge	56%	28%	$<.001$
MME per hospital day > 10	80%	52%	$<.001$

Aims

- Accurate predicting $\rightarrow$ improving sensitivity and specificity for imbalanced outcome
- Using and comparing methods of probability cutpoints and sampling

Methods

- Create sampled datasets
- Run model on sampled data
- Get predicted probabilites on the test data
- Optimize probability cutoff for outcome

Model

- Roughly 2/3 temporal split of data to get train and test set
- Cross validated lasso regression

Lasso

- Lasso:
  - Performs variable selection
  - Shrinks estimates to 0

$$\hat{\beta}_{lasso} = argmin \sum_{j=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq t$ where t is the tuning parameter.

Cross Validation

- Cross validattion:
  - Find the best “tuning measure” for model selection which determines amount of shrinkage of estimates
  - Split data into k parts and then train on each of those except one you validate against
  - Then pick the tuning measure that minimizes error

Advantages and Disadvantages

- Advantages:
  - Lower variance of the predicted values?
  - More accurate predictions
  - Reduces the number of predictors
- Disadvantages:
  - No interpretation of predictor coefficients
  - No standard errors out of the model
  - Biased coefficients

ROC & Cutoff Probablities

ROC (with pROC package):

- ROC curve plots sensitivity vs specificity
- Top left corner is ideal
- Youden Index is the furthest upper left corner or “max”

ROC Curve

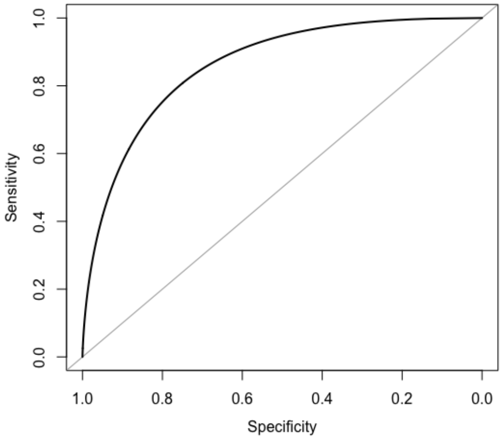

Figure 1: ROC Curve

Confusion Matrix

- show matrix
- show equations to calculate sens, spec, npv, ppv

First approach

No Sampling, Optimze Cut-off:

- Use original unsampled data and get predictions off the lasso model
  - Predictions return probability between 0 and 1 for each observation
- Use 0.5 standard probability cutoff to compare
- Find "best" probability cutoff
  - Youden Index

Second Approach

Sampling:

- Create sampled data sets that are balanced
  - Down sample
    - under-sample majority to equal minority
  - Up sample
    - over-sample minority to equal majority
  - SMOTE
    - Synthetic Minority Over-sampling Technique
- Predict and use Youden Index as cutoff

Results

Data	Threshold	Specificity	Sensitivity	NPV	PPV	Accuracy	AUC
Unsampled 0.5	0.5	99	8	96	35	96	86
Unsampled	0.043	73	85	99	12	73	86
Down sampled	0.401	73	85	99	12	74	86
Up sampled	0.399	74	85	99	12	74	87
SMOTE	0.472	84	74	99	17	84	86

ROC Plot

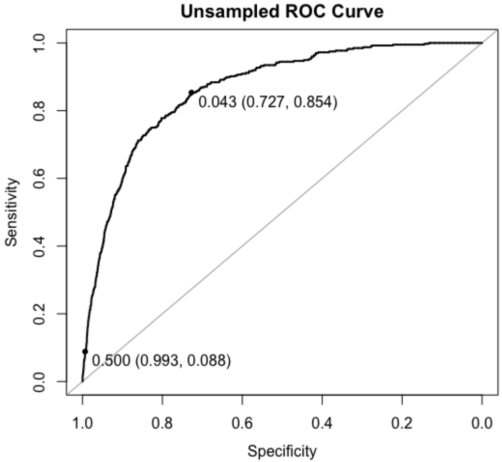

Figure 2: ROC for Original Data: Younden and 0.5 cutoffs

ROC Plot

plot with all of them on top of each other and show how similar they are?

Conclusions Thus Far

- Seeing similar results for both methods
- Depening on situation the clinician may like different sensitivity/specificity
- Some may want to be more conservative, others may not
    - Example: cancer patients in a significant pain

Moving Forward

- Simulation of different percentages for rare outcomes
  - which method performs the best for 5% 10% 50%
  - plogis(y, 1, int + b1 x age + b2 x op_receipt + . . .)
  - where int <- -1.5 (vary this) b1 <- x (can use estimates from Susan's paper)
  - we get a y (the only thing we need to generate, not the data)
  - then use lasso with y and xmatrix and predict to get confusion matrix
  - table 3 will be prevelance (5, 20, 50, etc maybe 10 values), and youden vs sampling with the different sens, spec
- Try different sampling other than defaults for each method
- Bagging (bootstrap aggregating)
- bootstrap aggregate the coefficients and get bootstrap CI (loop through getting new sample, saving coefficients, get mean and sd across 1000 boot samples)

More Moving Forward

- look up bagging and stepwise selection
- haven't seen much on bagging and down sampling– look that up, if not that'll be interesting
- check to see if there's a package to do bagging with lasso
- feed final average model with test set
- package SparseLearner? or Predict.bagging
- because when we down sample we only get one subset

Timeline

Defend March 15?

Questions?

Questions?

Notes

show smooth spline of age and probability of COT and show it has a curve to it and why we added the quadratic age smooth.spline with 3 degrees of freedom - Where would it make sense to include this???

accumulate other papers in endnote

(Check out most recent manuscript from paper an citations)

Notes 11.14.17

look at logistic and lasso results and compare (probably not necessary to include)

See if there are any other ways to choose a threshold

Might want to come up with another methodlogical advancement

DO THIS!! Just notes and what lit to look into Abstract, Background, methods, results, conclusions, acknowledgments, tables, figures, think about what's table 1 Be good to bring to meeting with Elizabeth to give her an idea of what we're working on!!!

Just showed variables for where there is a really big split between outcomes could use those and then maybe a couple others

mention TRIPOD - mention this, about being transparent, read paper