Proposal Outline

Alyssa Forber

October 16, 2017

Background/Problem

(Check out most recent manuscript from paper an citations) call it progression to opioid therapy

Opioid information

- explain issue of opioid addiction
- rare outcome
- whether or not to prescribe to patients if likely to become addicted

Intro

The United States is facing an unprecedented opioid epidemic. According to data from the 2015 National Survey of Drug Use and Health, over 2 million people had a prescription opioid use disorder.[1]

This is particularly important in the hospital where opioids are commonly prescribed for pain.[10] Opioid receipt at hospital discharge has been shown to be associated with an increased risk of chronic opioid use.[11]

Predictive tools to identify hospitalized patients at risk for future COT may have clinical utility to improve hospital-based pain management with a focus on limiting opioid prescribing when non-opioid analgesics, or other non-pharmaceutical options, may be effective for pain control.

Data

Describe dataset (comes from Denver health, more info in paper)

- show a table1 of the data
- number of subjects
- outcome
- variables in dataset

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.

The study had a binary outcome of chronic opioid therapy (COT) one year following the index discharge. We defined COT as receipt of $>=$ 90-day supply of opioids with $<$ 30-day gap in supply over a 180-day period or receipt of $>=$ 10 opioid prescriptions over one year. This is a rare outcome. 27705 in dataset, 1457 with outcome, about 5%

There are 50 variables, which do I mention? And then we narrowed to start with 35 Demographics and potential predictors Data dictionary I could look at?

Aims?

- accurate predicting, better sens and spec for unbalanced outcome
- using and comparing methods of cutpoints and sampling

Methods

describe analysis approach

Model

- Roughly 2/3 temporal split of data to get train and test set
- Cross validated lasso regression
- Lasso:
- Cross validattion:
  - Fet the best "tuning measure" for model selection
  - 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?

First approach

Cut-off Probabilities:

- 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
- Find "best" probability cutoff
  - Youden Index

ROC & Cutoff Probablities

ROC (with pROC package):

- explain ROC curves and youden
- show an example ROC curve
- Top left is ideal
- Or choose based off certain sensitivity or specificity

Confusion Matrix

Give example of confusion matrix and equations for calculating sens, spec, etc??

ROC Plot

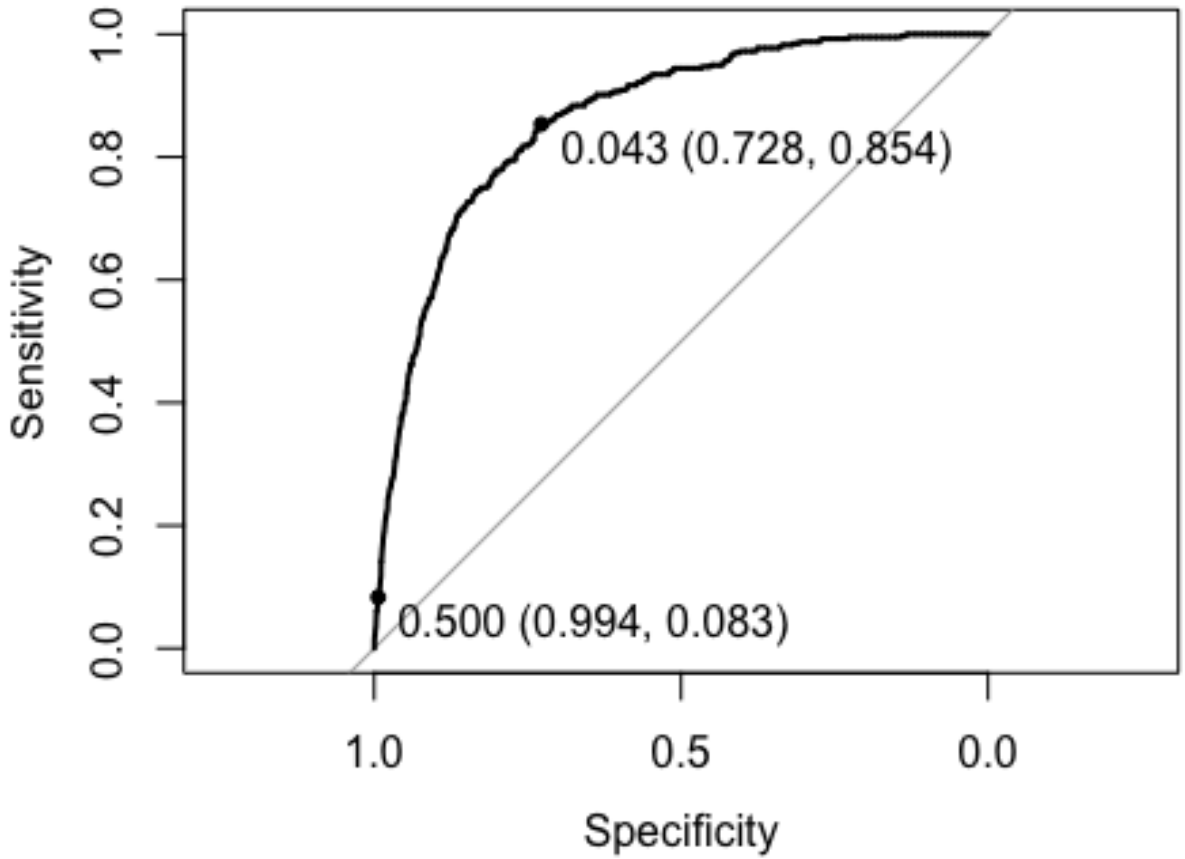

Figure 1: ROC for Original Data: Younden and 0.5 cutoffs

Second Approach

Sampling:

- Create sampled data sets that are balanced
  - Down sample
  - Up sample
  - SMOTE
    - ★ Synthetic Minority Over-sampling Technique
- Predict with Youden cutoff

Preliminary Results

Explain prelim results (sens, spec, accuracy, AUC, npv, ppv):

Explain meaning of sens, spec, npv, and ppv???

Table of–

- non-sampled with 0.5 cutoff
- non-sample with youden cutoff
- up sample with youden cutoff
- down sample with youden cutoff
- smote sample with youden cutoff
- down sample with bagging (DONT HAVE THIS)

Results Table

	Threshold	Specificity	Sensitivity	NPV	PPC	Accuracy	AUC
Unsampled 0.5	0.5	99.4	8.3	96.3	35.1	95.7	86.4
Unsampled Youden	0.043	72.8	85.4	99.2	11.7	73.3	86.4
Down Sampled	0.401	73	85.1	99.1	11.7	73.5	86.4
Up Sampled	0.399	73.8	85.4	99.2	12.1	74.3	86.5
SMOTE	0.472	84.1	74.2	98.7	16.5	83.7	86.4

Figure 2: Results by Model

Discussion of Results

- 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
  - When you could run model without sampling or changing cutoffs (though sampling does allow a more parismonious model)
- 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) With this you may get 12 var with one stepwise and 13 with another

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
- maybe try lasso with cross validation (cv.lasso)
- lasso has been shown to be better at selecting a model than stepwise
- easy to save all the coefficients and bootstrap
- 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
- she'll send surgical infections code

Questions

Should I show code in this?

Am I giving the talk as if for an audience of statisticians or clinicians or mixed?

What more should be expanded on?

Any figures or tables that would be helpful

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???

NOTES FROM 10/31

SCHEDULE!!! schedule between dec 4 and 15 send susan a note (she'll be hardest) ask what days and times are best could send out a doodle poll then send out her availablity

accumulate other papers in endnote

Thoughts

I think I'm confused on cv and bagging with lasso I need a better understanding of lasso first

so cv splits my sample into k subsamples and runs the model for k-1 of them and compares to the 1 left? this isn't making sense in the scheme of already making my training and test sets with the temporal split– it seems like it doesn't make sense to do both?

and then bagging is training many models on resampled data and then take their average to get an averaged model. this makes sense, it's how I get my model with resampling my training data and then I fit it with the test data

SoI don't understand cv well and I definitely don't understand doing both together

Thoughts cont.

I understand cv better now maybe– I use cv.glmnet to find my minimum lambda. I know that is the best. I'm not sure what it is, but it optimizes something. Then I use that min for glmnet and predict on that.