Identification of surgical site infections using electronic health record data and methodologies for imbalanced learning

Katie Colborn

Department of Biostatistics and Informatics
University of Colorado Anschutz Medical Campus

April 18, 2018
Background on ACS NSQIP
Complications tracked by NSQIP
Published models for infection identification
A model for identifying surgical site infections at UCH
Imbalanced learning
Results & Conclusions
Future directions
The American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) began in 2005 with the goal of assisting hospitals with identifying and preventing surgical complications.

Each participating hospital assigns a surgical clinical nurse reviewer to collect preoperative through 30-day postoperative data on a sample of surgical patients in order to risk adjust postoperative complications so that they can be compared across participating hospitals.

At large volume hospitals these samples might represent only 10-15% of all surgical cases.

ACS NSQIP data are considered gold standard for accurate identification and comparison of postoperative complications.
Chart review is time-consuming and costly and, as a result, cannot easily be scaled up to cover all surgical patients.

18 complications tracked: infectious (6), cardiac/transfusion (3), pulmonary (4), venous thromboembolic (2), renal (2), neurological (1).

Most common complication is postoperative infection (7%).

Postoperative infections are classified as: surgical site infections (SSIs), urinary tract infections (UTIs), pneumonia, and sepsis.
Researchers have developed models for identifying surgical infections through supervised learning approaches with mixed success.

Previous retrospective cohort studies have identified SSI electronically from medical records using multivariable regression models with around 64% sensitivity and greater than 90% specificity.

Most other types of infections were identified correctly between 40-75% of the time, and almost all achieved greater than 90% specificity.

The low sensitivity achieved in these previous studies was likely due to 1) the imbalance in the data (i.e., low prevalence), or 2) lack of important independent variables.
Objective of this study

Using supervised learning, develop a model for identifying SSIs within 30 days after surgery using EHR data linked to the NSQIP patient outcomes.
Methods

- Conducted literature search to determine which variables to include
- Generated lists of CPT codes, ICD-9 codes and medications sorted by frequency that were observed in our data between 3-30 days post-operation
- Two physicians each went through the lists separately and marked the codes that represented potential treatment or identification of SSI
- Those that they both agreed on were included as independent variables
This generated a list of 134 variables.
The data were split into training (before 2016, n=5191) and test sets (2016, n=1646).
Only 3% of patients had an SSI.
We chose to fit a generalized linear model (GLM) with a lasso penalty.

Penalty was chosen using 10-fold cross-validation.

Several techniques were explored to handle the imbalance.
Imbalanced learning

- When one class greatly outnumbers the other, the outcome is said to be imbalanced.
- Models fit to imbalanced data tend to produce very low sensitivity (when the class of interest is the minority).
- Methodologies for improving sensitivity address this either through post-processing model predictions or by altering the training data.
The most straightforward approach to maximizing sensitivity and specificity is to adjust the probability threshold for classification using the ROC curve. This is done after the model has been tuned and fit. After fitting a GLM with lasso penalty to the training data, and using Youden’s $J$ statistic to choose the threshold, we get the following results:

- 0.036 (0.900, 0.803)
- 0.500 (0.999, 0.230)
Changing the cutoff can only move samples up and down rows of the confusion matrix, not from the off-diagonals to the diagonals.

<table>
<thead>
<tr>
<th></th>
<th>0.5 cutoff</th>
<th>0.036 cutoff</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no SSI</td>
<td>SSI</td>
</tr>
<tr>
<td>no SSI</td>
<td>1583</td>
<td>2</td>
</tr>
<tr>
<td>SSI</td>
<td>47</td>
<td>14</td>
</tr>
</tbody>
</table>
Sampling methods

- When there is *a priori* knowledge of class imbalance, sampling methods can be used to select a training set such that the response rates are equal.

- However, it is important to still select the test set to reflect reality (i.e., it should reflect the imbalance in the response variable).

- For the SSI problem, we used a temporal split, where the training data consisted of patients operated between 2013-2015 and the test set consisted of patients operated in 2016.

- SSI prevalence was 3.3% in the training set and 3.7% in the test set.
When using existing data, such as an EHR, post hoc approaches are used:
- *Down-sampling* or *under-sampling*
- *Up-sampling* or *over-sampling*
- Synthetic minority over-sampling technique (SMOTE; Chawla, 2002)
Sampling methods continued

- **Up-sampling**: sampling the minority class with replacement until each class has approximately the same number
- **Down-sampling**: select a sample from the majority class equal in size to the minority class
SMOTE uses both up-sampling and down-sampling

- It has 3 operational parameters: 1) amount of up-sampling, 2) amount of down-sampling, and 3) number of neighbors used to impute new cases

- To up-sample, a data point is randomly selected from the minority class and its $K$-nearest neighbors are determined

- The new synthetic data point is a random combination of the predictors of the randomly selected data point and its neighbors
Take the difference between the feature vector under consideration and its nearest neighbor.

If $k = 5$ nearest neighbors, then 1 is randomly selected from the 5 for 100\% over-sampling; if 200\% is desired, choose two randomly for each case.

Multiply this difference by a random number between 0 and 1, then add it to the feature vector under consideration.
<table>
<thead>
<tr>
<th>Model</th>
<th>t</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Accuracy</th>
<th>NPV</th>
<th>PPV</th>
<th>AUC</th>
<th>No. variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full: .05</td>
<td>0.5</td>
<td>0.999</td>
<td>0.23</td>
<td>0.97</td>
<td>0.97</td>
<td>0.88</td>
<td>0.89</td>
<td>35</td>
</tr>
<tr>
<td>Full: t</td>
<td>0.036</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.99</td>
<td>0.24</td>
<td>0.89</td>
<td>35</td>
</tr>
<tr>
<td>Under-sampled: .05</td>
<td>0.5</td>
<td>0.88</td>
<td>0.8</td>
<td>0.87</td>
<td>0.99</td>
<td>0.2</td>
<td>0.89</td>
<td>20</td>
</tr>
<tr>
<td>Under-sampled: t</td>
<td>0.48</td>
<td>0.87</td>
<td>0.82</td>
<td>0.87</td>
<td>0.99</td>
<td>0.2</td>
<td>0.89</td>
<td>20</td>
</tr>
<tr>
<td>Over-sampled: .05</td>
<td>0.5</td>
<td>0.88</td>
<td>0.8</td>
<td>0.87</td>
<td>0.99</td>
<td>0.2</td>
<td>0.89</td>
<td>123</td>
</tr>
<tr>
<td>Over-sampled: t</td>
<td>0.45</td>
<td>0.91</td>
<td>0.79</td>
<td>0.9</td>
<td>0.99</td>
<td>0.24</td>
<td>0.89</td>
<td>123</td>
</tr>
<tr>
<td>SMOTE: .05</td>
<td>0.5</td>
<td>0.89</td>
<td>0.75</td>
<td>0.89</td>
<td>0.99</td>
<td>0.21</td>
<td>0.88</td>
<td>88</td>
</tr>
<tr>
<td>SMOTE: t</td>
<td>0.15</td>
<td>0.79</td>
<td>0.89</td>
<td>0.8</td>
<td>0.99</td>
<td>0.14</td>
<td>0.88</td>
<td>88</td>
</tr>
</tbody>
</table>
## Results: comparison to models in the literature

<table>
<thead>
<tr>
<th>Model</th>
<th>Branch-Elliman et al.</th>
<th>Comprehensive Model</th>
<th>Combination Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.84</td>
<td>0.9</td>
<td>0.89</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.82</td>
<td>0.8</td>
<td>0.82</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.84</td>
<td>0.9</td>
<td>0.88</td>
</tr>
<tr>
<td>NPV</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>PPV</td>
<td>0.16</td>
<td>0.24</td>
<td>0.22</td>
</tr>
<tr>
<td>False negatives</td>
<td>11</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>False positives</td>
<td>255</td>
<td>157</td>
<td>181</td>
</tr>
<tr>
<td>AUC</td>
<td>0.86</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>Number of variables selected</td>
<td>3</td>
<td>35</td>
<td>20</td>
</tr>
</tbody>
</table>
Model: full training and threshold method

Predicted Prob

Frequency

Misclassification Error

log(Lambda)

Misclassification ... 129 125 124 113 98 84 69 46 28 15 9 4 3 2

Katie Colborn

Informatics Symposium 20/23
Conclusions

- In these data, the threshold method and sampling perform similarly in the hold-out set.
- There was a slight edge for using full training data and threshold method and not combining covariates into groups.
- Although 80% sensitivity is acceptable, perhaps the model could be improved with the inclusion of claims data.
- I did not show the results in this talk, but other algorithms (gradient boosting machine, random forest, logistic regression) all performed similarly (but slightly worse).
Other types of infections (UTI, sepsis, pneumonia) will be explored similarly, thanks to a recent AHRQ R03 award.

Ideally, these models can be used to assess the impact of large-scale interventions.
Acknowledgments

- This work was funded by a D2V project grant
- Thanks to Health Data Compass for supplying data and linking to the NSQIP patients
- Thank you to my colleagues:
  - Michael Bronsert, Ph.D., ACCORDS
  - Karl Hammermeister, M.D., ACCORDS
  - Bill Henderson, Ph.D., Department of Biostatistics and Informatics and ACCORDS
  - Rob Meguid, M.D., Department of Surgery and ACCORDS