Using Oversampling in Machine Learning Models for Patient Risk Stratification for Acute Lower Gastrointestinal Bleeding

Jonathan JC Ho, MD[1], George M. Hanna III, MD[1], Amber Charoen, MD[2], Fadlallah Habr, MD[3]
[1]Department of Internal Medicine, Warren Alpert School of Medicine, Brown University, Providence, RI
[2]Department of Medicine, Division of Gastroenterology and Hepatology, Johns Hopkins University School of Medicine, Baltimore, MD
[3]Division of Gastroenterology, Warren Alpert School of Medicine, Brown University, Providence RI

Introduction

Lower gastrointestinal bleeding (LGIB) is a common cause of hospital admissions and can lead to hospital-based interventions that consume a significant amount of medical resources. However, only a minority of cases are high-risk and result in significant morbidity and mortality. We present an oversampling method to help with rebalancing for machine learning modeling for triaging in LGIB when there is significant imbalance between high risk (HR) and low risk (LR) patients.

Method

From retrospective data, hemodynamically stable patients with suspected LGIB were labeled into HR or LR groups (Figure 1). Risk factors associated with LGIB (e.g. age, sex, blood pressure, hemoglobin) were included as predictors. The dataset was divided into 80% for training and 20% for testing. Two machine learning models (stepwise logistic regression and decision trees) were applied to the training data to create predictive models. Then, the training and testing performances were evaluated using standard performance metrics (e.g. sensitivity, specificity, and F1).

Conclusion

Logistic regression did not perform as well as decision trees in training; however, it can generalize better to unseen data. Obtaining more HR cases can reduce the overfitting issue and provide a more accurate predictive model.

Results

Overall 1414 records were reviewed. General characteristics are demonstrated in **Tables 1 and 2**. There were 69 HR patients and 1345 LR patients. Among the included factors, age, blood pressure, pulse, BUN, Hb, INR, quartile of transfusions, and being on antiplatelet agents were statistically different between the 2 risk groups. **Table 3** shows the statistical results for the training and testing phases. Logistic regression model scores were normalized to 10, and cut-offs were plotted on an ROC curve (**Figure 2**).

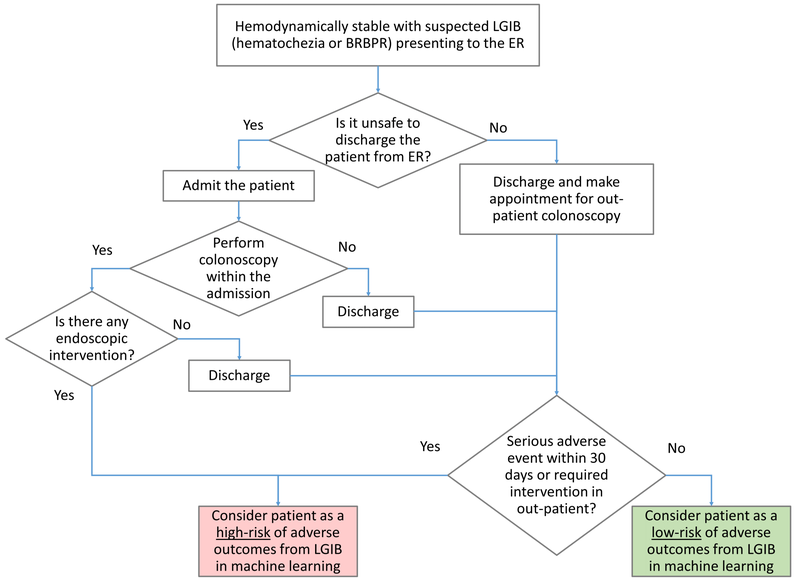

Figure 1. Algorithm for classifying low-risk and high-risk of LGIB. BRBPR = bright red blood per rectum, ER = emergency room, LGIB = lower gastrointestinal bleeding

	Training		Testing	
	Logistic Regression	Decision Trees	Logistic Regression	Decision Trees
Positive Class	586 (51.8%)	586 (51.8%)	12 (4.3%)	12 (4.3%)
Accuracy	0.7067	0.9452	0.7447	0.9433
Sensitivity/Recall	0.6911	0.9078	0.5833	0.0833
Specificity	0.7234	0.9853	0.7519	0.9815
Pos Pred Value/Precision	0.7284	0.9852	0.0946	0.1667
Neg Pred Value	0.6858	0.9088	0.9760	0.9601
F1	0.7093	0.9449	0.1628	0.1111

Table 3. Performance metrics of logistic regression compared to decision trees.

Table 1. General characteristics

Factors	Values
Age (years)	61.0 (44.0,76.0)
Sex (male)	690 (48.8%)
Alcohol use	656 (46.39%)
Drug use	232 (16.41%)
Blood pressure – systolic (SBP)	138.0 (124.0,154.0)
Blood pressure – diastolic (DBP)	78.0 (68.0,88.0)
Pulse	82.0 (73.0,92.0)
Anticoagulant use	198 (14.0%)
Antiplatelet use	61 (4.31%)
NSAID use	211 (14.92%)
Other procedures during admission	1404 (99.29%)

Table 2. General characteristics continued

Factors	Values
BUN	16.0 (12.0,22.0)
Hemoglobin (Hb)	13.2 (11.7,14.5)
Creatinine	0.83 (0.71,1.04)
Prothrombin time	11.9 (11.3,13.1)
INR	1.1 (1.0,1.2)
Platelets	227.0 (184.0,275.75)
WBC	7.9 (6.2,10.0)
Blood transfusion (Bt)	0.0 (0.0,0.0)
Last visit within 30 days	27 (1.91%)
High-risk of LGIB	69 (4.88%)

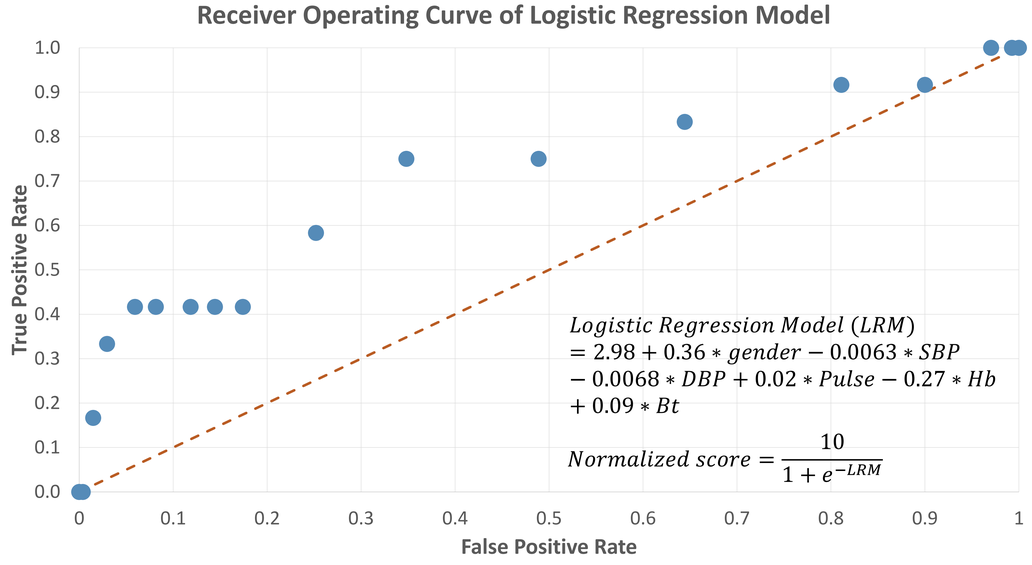

Figure 2. Receiver operative curve of the logistic regression model on testing data. The area under the curve is 0.754.