Title: Adverse outcome prediction of neurosurgical cases

Background: Predicting adverse outcomes is an essential part of risk management in surgery, where a rapid and accurate prediction would help inform decision-making and improve resource allocation. While the American College of Surgeons (ACS) has established an online regression-model based surgical risk score calculator to predict postoperative complications based on preoperative risk factors for different types of procedures, the prediction accuracy could potentially be further improved with machine learning models capable of leveraging more complex relationships between predictors and targets.

Research Questions: The study focuses specifically on neurosurgical outcome predictions and attempts to establish machine learning models that use preoperative risk factors and other basic patient characteristics to predict the occurrences of 30-day postoperative mortality and other adverse outcomes**.**

Predictors: Demographic information: sex, race, ethnicity (hispanic/not hispanic), and age. Preoperative risk factors: BMI class, functional status, American Society of Anesthesia (ASA) class, steroid use for chronic condition, ascites within 30 days prior to surgery, systemic sepsis within 48 hours of surgery, >10% weight loss in last 6 months, current smoker, ventilator dependent, bleeding disorder, disseminated cancer, diabetes, hypertension requiring medication, previous cardiac event, congestive heart failure in 30 days before surgery, dyspnea, history of severe chronic obstructive pulmonary disease (COPD), dialysis, acute renal failure, wound class, and emergency case.

Outcome: The 30-day postoperative adverse outcomes are categorized as [1]:

a) mortality (value = 3);
b) severe adverse events: prolonged intubation of 48 hours or more, return to the operating room, unplanned reintubation, sepsis, deep venous thrombosis/pulmonary embolism, coma, stroke, cardiac arrest, septic shock, surgical site or deep infection, and acute renal failure (value = 2);
c) minor adverse events: perioperative blood transfusion, urinary tract infection, pneumonia, progressive renal insufficiency, and wound disruptions (value = 1);
d) no adverse events (value = 0).

Method: The study is based on a subset of the ACS-NSQIP Dataset, which contains 144 variables with a total of 162, 945 entries from 2013. The dataset was given in csv format and cleaned using python. After dropping cases with unknown ages, cases with neurosurgery surgeon speciality were selected. Among the 144 variables, 4 encoded demographic information and 21 encoded preoperative risk factors, resulting in 25 predictors in total. The body mass index (BMI) class predictor was calculated from height and weight (missing values imputed with means) and categorized as underweight if < 18.5, normal weight if between 18.5 and 24.9, overweight if between 24.9 and 29.9, and obese if above 29.9 [2]. The outcome variable was determined from 5 variables that encoded minor 30-day postoperative adverse outcomes, 13 variables that encoded severe adverse outcomes, and the year of death variable that encoded death based on the aforementioned outcome section. No occurrence of these outcomes would be categorized as no adverse outcome for a case.
Upon examining predictor distributions (Table 1), the variable ascites had highly imbalanced distribution and was thus removed. As all variables except for age were ordinal, the Chi-square test of independence was also conducted to remove predictors largely unrelated to the outcome (Table 2).

With the final list of variables, supervised learning models were subsequently constructed, tuned, and evaluated in python. The following classifiers were used: Decision Tree, Decision Tree-based Bagging, Random Forest, Adaptive Boost (AdaBoost), Gradient Boost, and Extreme Gradient Boost (XGBoost). The data was split into 75% for train/cross-validation and 25% for test. Grid search with stratified 10-fold cross-validation was used to tune model hyperparameters to attain the highest accuracy based on the predefined hyperparameter grids (Table 3). Accuracy, the average of one-vs-rest area under the receiver operating characteristic curve (AUC), and support-weighted F1 scores were used to evaluate model performance.

Results: The dataset consists of 14,523 cases, among which 1,114 (7.67%) resulted in minor adverse outcomes, 586 (4.03%) resulted in severe adverse outcomes, 93 (0.64%) resulted in mortality 30-day postoperatively, and the rest no adverse outcomes (87.66%). 3.16% of the cases were labeled as emergency, implying the patients' conditions were aggravated by preoperative delay and could quickly deteriorate. Mean age was 58.5 y.o., ranging from 18 to 89 y.o.. 48.61% of the patients were female and 51.39% were male. The majority of the cases were white patients (82.10%). All predictors except for anesthetic type were found to be having significant associations with the adverse outcome variable (p-*value* < 0.05).

The dataset was stratified into 75% train/cross-validation (10,892 cases) and 25% test (3,631 cases) reproducibly with a random seed of 42. After selecting the best model set-up, learning curves, confusion matrices (Figure 1), and evaluation metrics were obtained for each model (Table 3). The models performed similarly on the test set in terms of accuracy and F1 scores, with AdaBoost classifier having the highest accuracy of 0.898 and the highest F1 score of 0.879. The highest AUC of 0.810 was achieved by XGBoost and followed closely by DT-Bagging classifier's 0.802, indicating good discriminative power of the model. On the other hand, Decision Tree and Gradient Boost performed poorly in terms of AUC, with values of 0.679 and 0.675 respectively.

However, it is worth noting that there exists a huge imbalance among the outcome classes, with the majority class of no adverse outcome taking up to more than 87% of the dataset. Without implementing strategies like resampling or cost-sensitive learning to address such a problem, the models all performed poorly on classifying the minority classes. The misclassification of the minority classes as the majority class was reflected in the confusion matrices (Figure 1). Additionally, the still rising trends of the test accuracy in the majority of learning curves as the sample size reached maximum indicated that the model's performance could be further improved with more data, although not necessarily in correctly classifying the minority classes.

Discussion:

In this study, several ensemble machine learning models and their base model decision tree were evaluated on the prediction of occurrences of 30-day postoperative mortality and other adverse outcomes in neurosurgical operations using preoperative risk factors and other basic patient characteristics. Based on the evaluation metrics, AdaBoost classifier is the one of the best models in terms of accuracy (0.898), F1 score (0.879), and a decent AUC (0.767), alongside XGBoost and DT-based bagging classifiers, which have slightly lower accuracy and F1 score but AUC above 0.8. However, as previously mentioned, these models did not have clinical applicability as they failed to correctly predict more than half of the occurrences of adverse events.

Given the risk predictions of the minority classes, which indicate either morbidity or mortality, are the most relevant to surgeons and patient families' shared decision-making and informed consent, the

model evaluation should take into consideration the class distribution imbalance and error cost difference. Therefore, in terms of evaluation metrics, accuracy might not be the ideal one for model hyperparameter tuning and test performance evaluation, as it is susceptible to inflation by the majority class. Instead, weighted accuracy could be calculated to adjust for the different prediction error costs; as well as class-specific Brier score and AUC to measure both model calibration and discrimination on each type of adverse outcomes as in the original ACS risk score calculator publication [3]. In terms of model training, reassigning class weights and upward sampling the minority classes could compensate for the class imbalance. Although both methods have been warned against due to strong distortions of probability estimates and miscalibrations observed in certain risk prediction models [4].

The most intuitive way to improve model performance is perhaps to increase the dataset size. The original dataset used to develop the risk calculator was approximately 10 times the size of the dataset used in this study, leading to much better model discrimination (AUC > 0.8) than in this study. The learning curves also showed signs of underfitting. Given the small event fraction and high expected prediction performance of the adverse outcomes, more data from various hospital sites should be used to develop such a kind of risk prediction model to ensure clinical utility and generalizability.

Reference:

[1] Lukasiewicz, A. M., Grant, R. A., Basques, B. A., Webb, M. L., Samuel, A. M., & Grauer, J. N. (2016). Patient factors associated with 30-day morbidity, mortality, and length of stay after surgery for subdural hematoma: a study of the American College of Surgeons National Surgical Quality Improvement Program. *Journal of neurosurgery, 124*(3), 760–766. https://doi.org/10.3171/2015.2.JNS142721

[2] Centers for Disease Control and Prevention. (2022, June 3). *All about adult BMI.* https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

[3] Bilimoria, K. Y., Liu, Y., Paruch, J. L., Zhou, L., Kmiecik, T. E., Ko, C. Y., & Cohen, M. E. (2013). Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *Journal of the American College of Surgeons, 217*(5), 833–42.e423. https://doi.org/10.1016/j.jamcollsurg.2013.07.385

[4] van den Goorbergh, R., van Smeden, M., Timmerman, D., & Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association : JAMIA, 29*(9), 1525–1534. https://doi.org/10.1093/jamia/ocac093

Table 1: Key variable statistics

	Height	Weight	Age
mean	66.54	190.84	58.48
std	4.14	47.41	13.46
min	48.00	70.00	18.00
max	86.00	572.00	89.00

Emergency case	No	14064
	Yes	459
Systemic Sepsis	None	14334
	SIRS	130
	Sepsis	50
	Septic Shock	9
Blood transfusion	No	14484
	Yes	39
Bleeding disorder	No	14271
	Yes	252
Weight loss	No	14443
	Yes	80
Steroid use	No	13862
	Yes	661
Wound infection	No	14417
	Yes	106
Disseminate cancer	No	14254
	Yes	269
Dialysis	No	14470
	Yes	53
Preop renal failure	No	14499
	Yes	24
Hypertension medication	No	6906
	Yes	7617
History of CHF	No	14464
	Yes	59
Ascites	No	14519
	Yes	4
History of severe COPD	No	13804
	Yes	719

BMI_class	Normal weight	2822
	Obese	6790
	Overweight	4783
	Underweight	128
Ethnicity (Hispanic/ not hispanic)	No	12970
	Unknown	820
	Yes	733
Race	American Indian or Alaska Native	115
	Asian	236
	Black or African American	1237
	Native Hawaiian or Pacific Islander	31
	Unknown/Not Reported	980
	White	11924
Sex	female	7059
	male	7464

Ventilator use	No	14473
	Yes	50
Functional status	Independent	13965
	Partially Dependent	375
	Totally Dependent	43
	Unknown	140
Dyspnea	AT REST	34
	MODERATE EXERTION	713
	No	13776
Current smoker	No	10977
	Yes	3546
Diabetes	INSULIN	820
	NO	11961
	NON-INSULIN	1742
Anesthesia type	Epidural	19
	General	14453
	Local	2
	MAC/IV Sedation	23
	Other	7
	Spinal	18
	Unknown	1

Adverse outcome	None	12730
	Minor	1114
	Severe	586
	Mortality	93

Table 2: Chi-square test of independence between categorical predictors and outcome

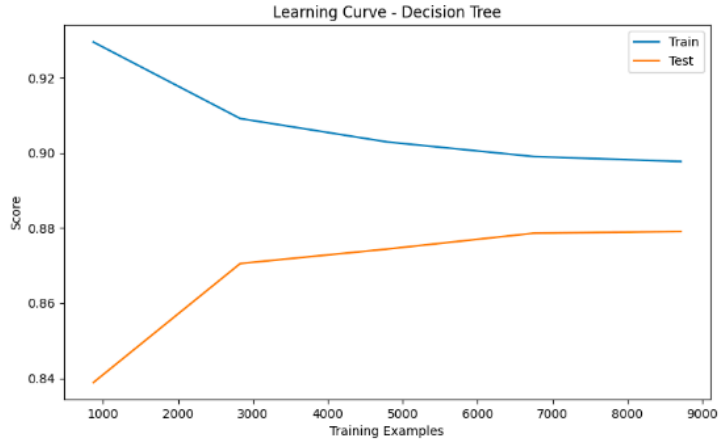
Predictor	**Chi-Square**	**p-value**
Hypertension medication	65.1	< 0.0001
Blood transfusion	262	< 0.0001
Bleeding disorder	49.8	< 0.0001
Weight loss	81.37	< 0.0001
Steroid use	45.2	< 0.0001
Wound infection	3494	< 0.0001
Disseminate cancer	524	< 0.0001
Dialysis	205	< 0.0001
Preop Renal failure	62.5	< 0.0001
Systemic sepsis	6719	< 0.0001
Emergency case	382	< 0.0001
History of severe COPD	20.7	< 0.0001
Ventilator use	748	< 0.0001
Functional status	644	< 0.0001
Current smoker	31.6	< 0.0001
Diabetes	98.0	< 0.0001
BMI_class	43.2	< 0.0001
History of CHF	26.7	< 0.0001
Dyspnea	19.0	0.004
Ascites	10.2	0.017
Anesthesia	7.65	0.983

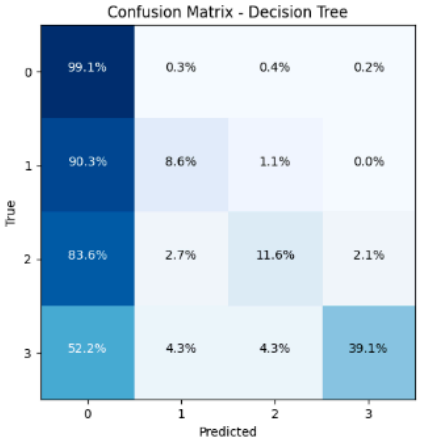
Table 3: Machine learning models and classification performance

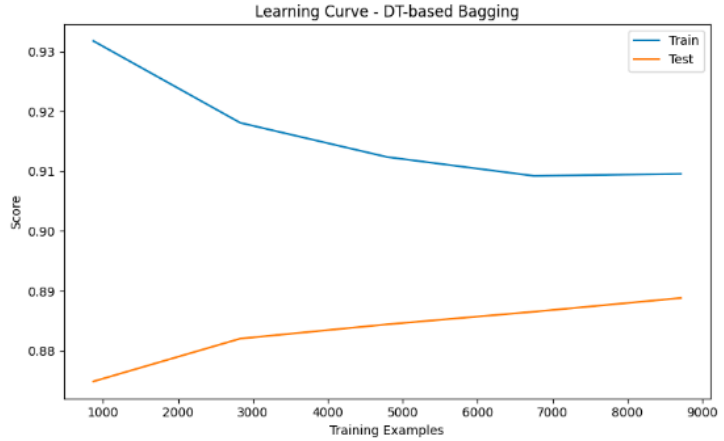
Model	Accuracy/F1 Score/AUC (10 fold stratified cross-validated)	Accuracy/F1 Score/AUC (Test)	Hyperparamter grid	Final set-up
DT-based Bagging	**0.888**/0.848/0.744	0.890/0.849/0.802	bootstrap: [True, False], bootstrap_features: [True, False], n_estimators: [5, 10, 15], max_samples: [0.6, 0.8, 1.0], estimator__max_depth: [1, 2, 3, 4, 5, 10, 20, 50], estimator__max_features: [0.6, 0.8, 1.0]	bootstrap=False, bootstrap_features=True, max_samples=0.8, n_estimators=50. Base estimator: DecisionTree: max_depth=20, max_features=0.8.
Random Forest	0.884/0.839/0.761	0.885/0.840/0.765	n_estimators = [10, 20, 50, 100, 150, 200] max_depth = [1,3,5,10,15] bootstrap = [True, False]	bootstrap=False, max_depth=15, n_estimators=150.
Decision Tree	0.882/0.842/0.625	0.883/0.845/0.679	criterion: ['gini', 'entropy'], max_depth: [5, 10, 15,20,30,40,None], min_samples_split: [2, 5, 10], min_samples_leaf: [1, 2, 4]	criterion='entropy', max_depth=10, min_samples_leaf=1, min_samples_split=2.
AdaBoost	**0.888**/**0.868**/0.700	**0.898**/**0.879**/0.767	n_estimators = [10,20,30, 50, 100] learning_rate= [0.0001, 0.01, 0.1, 1.0, 1.5] base_estimator__max_depth[1] = [1, 3, 5, 10, 20] base_estimator__max_features=[0.6, 0.8, 1.0]	n_estimators=60, Base estimator: DecisionTree: max_depth=70, max_features=0.01.
XGBoost	0.886/0.849/**0.771**	0.891/0.858/**0.810**	n_estimators = [100, 200, 500] max_depth = [2, 3, 5, 10] learning_rate=[0.1,0.5,1] min_child_weight=[1,2,3]	booster='gbtree', objective='multi:softprob' Base, learning_rate=0.1, n_estimators=110, estimator: max_depth=10, max_leaves=None, min_child_weight=2.
Gradient Boost	0.876/0.819/0.638	0.877/0.819/0.675	learning_rates = [1, 0.5,0.1] n_estimators = [100] max_depths = [2, 3, 5, 10,20,30] min_samples_split = np.linspace(0.1, 1.0, 3, endpoint=True) min_samples_leafs = np.linspace(0.1, 0.5, 3, endpoint=True)	learning_rate=1, max_depth=10, min_samples_leaf=0.1, min_samples_split=0.1.

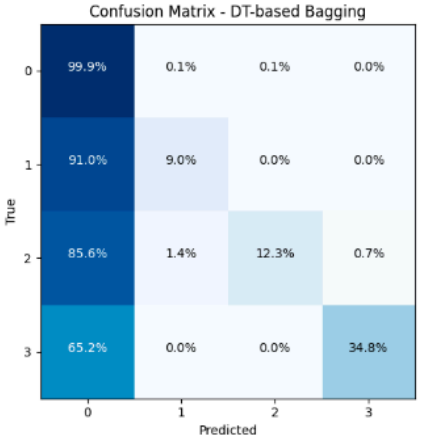
[1] In sklearn v1.2, the hyperparameters that tune the decision tree in the ensemble learners are named estimator_… instead of base_estimator_…

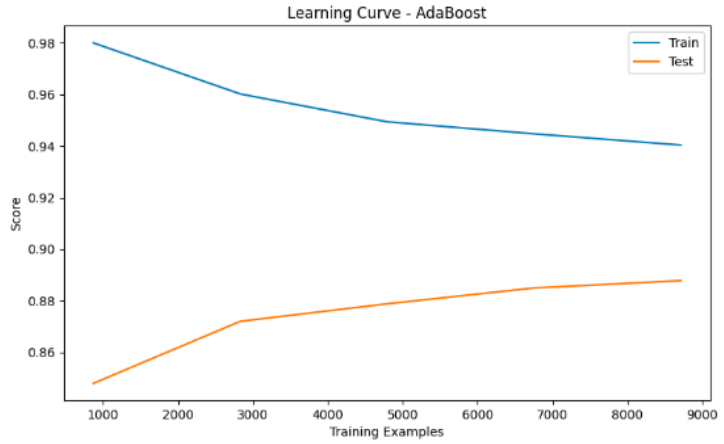
Figure 1: Confusion matrices and learning curves

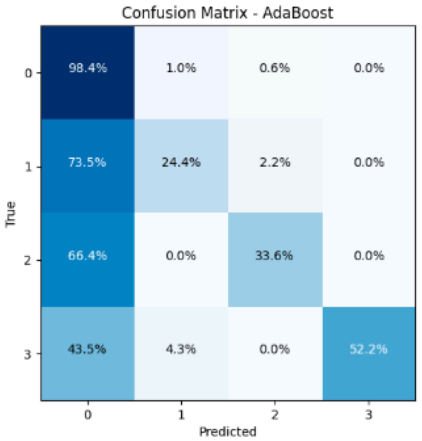


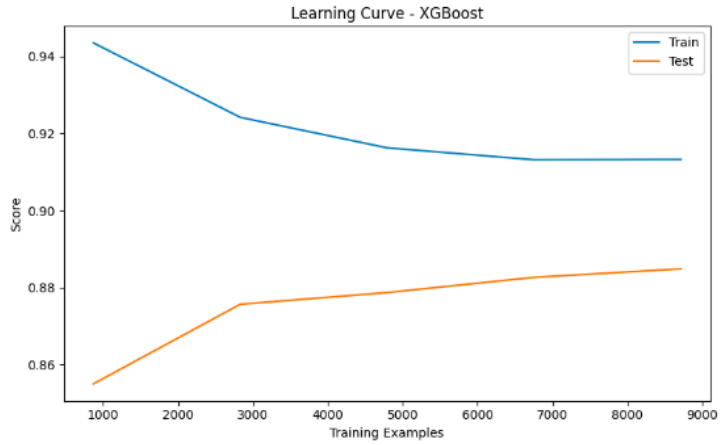


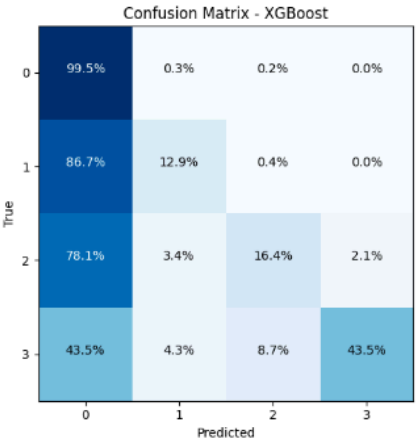


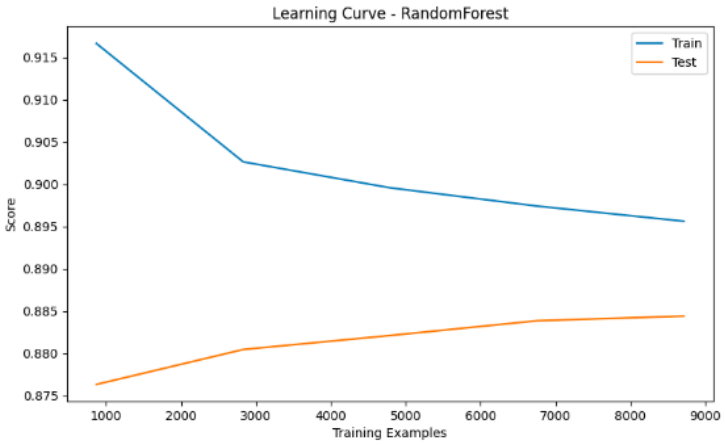


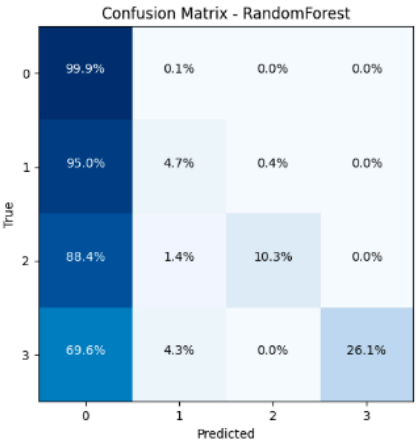


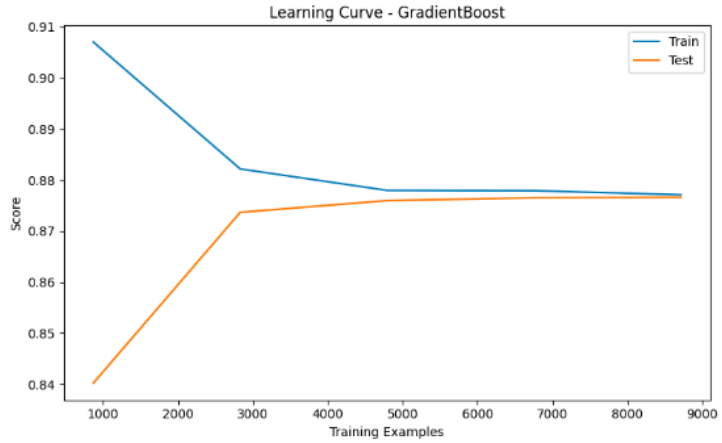


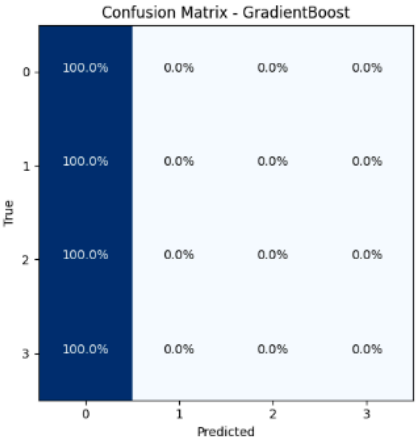

Learning Curve - XGBoost
Train
Test
Score
Training Examples


Confusion Matrix - XGBoost
True
Predicted
99.5% 0.3% 0.2% 0.0%
86.7% 12.9% 0.4% 0.0%
78.1% 3.4% 16.4% 2.1%
43.5% 4.3% 8.7% 43.5%


Learning Curve - RandomForest
Train
Test
Score
Training Examples


Confusion Matrix - RandomForest
True
Predicted
99.9% 0.1% 0.0% 0.0%
95.0% 4.7% 0.4% 0.0%
88.4% 1.4% 10.3% 0.0%
69.6% 4.3% 0.0% 26.1%


Learning Curve - GradientBoost
Train
Test
Score
Training Examples


Confusion Matrix - GradientBoost
True
Predicted
100.0% 0.0% 0.0% 0.0%
100.0% 0.0% 0.0% 0.0%
100.0% 0.0% 0.0% 0.0%
100.0% 0.0% 0.0% 0.0%