Extreme Gradient Boosted Classification and Regression Trees to Predict Outcomes in Patients with Monoclonal Gammopathies of Undetermined Significance

Dutta, Aheli (School: Nikola Tesla STEM High School)

Monoclonal gammopathy of undetermined significance (MGUS) is a premalignant condition at risk of progressing to multiple myeloma, a type of incurable blood cancer, and other type of plasma cell malignancies. Although patients may benefit from early treatment, current prognostic models have limited abilities to predict myeloma transformation. We hypothesized gradient boosted classification and regression tree models, a type of machine learning algorithm, could surpass conventional statistical models to predict myeloma transformation in MGUS patients. We trained classification and regression trees with extreme gradient boosting (XGBoost) using a cohort of 1400 MGUS patients. As a comparator, a logistic regression model was fit. Both models included the following features evaluated at the time of MGUS diagnosis: age, sex, hemoglobin, creatinine, and monoclonal protein serum concentrations. Both models also ranked feature importance. Model discrimination (the ability of the model to correctly rank patients with lower risk compared to those with higher risk) was assessed using the area under the receiver-operator curved (AUROC). Model calibration (ie, the agreement of the predicted risk with the observed proportions of patients experiencing the event) was assessed using calibration curves. The XGBoost model had an AUC of 0.61, compared to the 0.69 of the logistic regression model. The XGBoost model did not outperform conventional logistic regression in terms of discrimination and calibration. The low frequency of cases in the dataset may have impacted model training and prediction of myeloma transformation in MGUS patients.