Predicting Mortality in Patients with Extremity Soft Tissue Sarcoma: Development and External Validation of Machine Learning Models

Teja Yeramosu, Waleed Ahmad, Azhar Bashir, Gregory F Domson[1]
[1]VCU Medical Center

INTRODUCTION:

Soft-tissue sarcoma (STS) denotes a group of heterogeneous malignant tumors, with high malignancy grade and certain histologic subtypes precipitating a high-risk of death. As accurate prediction of adverse outcomes following diagnosis of STS can guide the triage of care services and shared decision-making, the objective of this study was to evaluate the utility of current machine learning methods in predicting cancer-related mortality in patients diagnosed with high-risk STS. Determining prognostic patient and tumor characteristics can help improve the identification of patients likely to benefit from adjuvant therapy as well as low-risk patients who may be spared the adverse effects of unnecessary treatment.

METHODS:

Demographic and clinicopathologic variables of high-risk STS patients in the Surveillance, Epidemiology, and End Results database from 2004-2017 were analyzed. Models were subsequently developed based on standard multivariable logistic regression (LR), support vector machine (SVM), naïve Bayes classifier (NB), boosted decision trees (BT), bootstrap forest (BF), and extreme gradient descent boosting (XGBoost). After splitting the data into training (75%) and validation (25%) data sets, the accuracies of these models were compared using the area under receiver operating characteristic curve (AUC). The model that performed best on the SEER testing data was again tested on our institutional patient data set to assess external validation.

RESULTS:

5,842 patients were diagnosed with STS, of which 1,793 (30.7%) experienced cancer-related mortality. The overall three most prevalent malignancies were leiomyosarcoma, undifferentiated pleomorphic sarcoma, and synovial sarcoma. Upon performing multivariable logistic regression analysis , the following variables were found to have a significant association: age (OR: 3.17 [2.22, 4.53], $p<0.0001$), grade (OR: 2.50 [2.09, 3.00], $p<0.0001$), T stage (OR: 2.71 [2.33, 3.15], $p<0.0001$), N stage (OR: 1.46 [1.01, 2.10], $p=0.0374$), M stage (OR: 4.82 [3.79, 6.14], $p<0.0001$), tumor size (OR: 2.49 [1.54, 4.04], $p=0.0003$), chemotherapy (OR: 1.38 [1.18, 1.61], $p=0.0001$), and surgery (OR: 0.50 [0.39, 0.64], $p<0.0001$). The three histological subtypes with the largest significant odds ratios were clear cell sarcoma (OR: 4.63 [2.39, 8.97], $p<0.0001$), malignant peripheral nerve sheath tumor (OR: 3.17 [2.36, 4.25], $p<0.0001$), and synovial sarcoma (OR: 2.89 [2.18, 3.84], $p<0.0001$). The LR, SGB, RF, SVM, and ANN models had AUCs of 0.76, 0.78, 0.84, 0.80, and 0.79 with the training dataset, respectively. The RF model further had AUC of 0.75 upon validation (Figure 1). The RF model identified tumor size as the most important variable when predicting mortality in patients with STS, followed by M stage, T stage, histological subtype, age, grade, surgical excision, chemotherapy, location of tumor, and race (Figure 2). Using the institutional dataset to externally validate the RF model developed using the SEER dataset yielded an AUC of 0.71.

DISCUSSION AND CONCLUSION:

To the best of our knowledge, this is the first study applying various machine learning techniques to predict mortality in patients with extremity STS. Among all five tested models, the RF yielded the highest AUC in the training dataset, followed by SVM, ANN, SGB, and finally, multivariable logistic regression, demonstrating good accuracy and predictability. The results from our RF model highlight specific demographic, clinicopathological, and treatment variables as being salient to its performance in predicting cancer-related mortality following STS diagnosis, larger size of tumor, greater M stage, greater T stage, histological subtype, specifically, clear cell sarcoma, malignant peripheral nerve sheath tumor, and synovial sarcoma, greater age, higher grade, lack of treatment with surgical excision or chemotherapy, location of tumor at the lower extremities, and non-white race. With the rising emphasis on value-based care, orthopaedic oncologists may incorporate models such as these to optimize treatment plans for their patients.

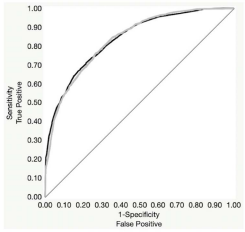

Figure 1. Receiver operating curve of Random Forest training model for patients with STS who did and did not experience cancer-related mortality. AUC reached 0.84.

AUC, area under the curve

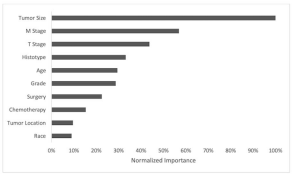

Figure 2. Normalized importance of demographic, clinicopathological, and treatment variables based on Random Forest model. Importance is the degree to which the model is dependent on the factor.