POSTER 16

Title: How is Missing Data Best Addressed When Developing and Externally Validating Machine Learning Models for Survival Prediction in Sarcoma?

Authors: Linus Lee, BE[1], Michael P. Fice, MD[1], Sarah C. Tepper, MD[1], Conor Jones, MD[1], Evan Klein, BS[1], Neil Buac, BS[1], Gayathri Vijayakumar, BS[1], Matthew W. Colman, MD[1], Steven Gitelis, MD[1], Alan T. Blank, MD, MS[1]

Author Affiliations:

1. Department of Orthopedic Surgery, Section of Orthopedic Oncology, Rush University Medical Center, Chicago, Illinois, USA; linus.h.lee@gmail.com, michael_p_fice@rush.edu, sarah_c_tepper@rush.edu, conor_m_jones@rush.edu, evan_d_klein@rush.edu, neilpatrick_l_buac@rush.edu, gayathri.vijayakumar8@gmail.com, matthew_w_colman@rush.edu, steven_gitelis@rush.edu, alan_blank@rush.edu

Background

National databases provide large cohorts which can be used to study rare diseases like sarcoma. While there is a large volume of cases, there is often a significant amount of missing data, and various methods of handling missing data have been shown to affect the results and conclusions in prior studies. The impact of missing data remains unclear in machine learning (ML) based models for survival prediction in sarcoma.

Methods

The Surveillance, Epidemiology, and End Results (SEER) database was queried for cases of undifferentiated pleomorphic sarcoma (UPS) and malignant fibrous histiocytoma (MFH) (n=4,299). Missing data were managed in four methods to create four distinct datasets: missForest data imputation (*full* dataset), exclusion of all cases with missing data (*excluded* dataset), missing variables treated as positive values (*positive* dataset), and missing variables treated as negative/null (*negative* dataset). These datasets were split into training:testing to create ML models for survival prediction and then models were externally validated on an institutional cohort of UPS patients. Model performance was assessed with c-statistics, Brier score loss, and calibration curves.

Results

The Logistic Regression (LR), Linear Support Vector Machine (LSVM), and Multi-layer Perceptron Neural Network (MLP) models all demonstrated good performance on 5-year survival prediction on internal validation with c-statistics ranging from 0.74 to 0.79. There was little variance among the models built off different datasets (maximum difference in c-statistic of 0.03 in each ML model. The maximum difference of c-statistic was 0.03 in each ML model. Brier score loss differed as much as 0.052 in the MLP model. All models were well calibrated. Model performance additionally did not differ greatly on external validation regardless of which dataset was used to train the model.

Conclusions

Missing data does not appear to greatly affect survival prediction in UPS when employing machine learning based models, which may be attributable to the iterative nature of machine learning. Nevertheless, clinicians should be cognizant of what data are being used to create ML models and interpret reported results with clinical judgment as ML models continue to evolve and eventually integrate into clinical practice.

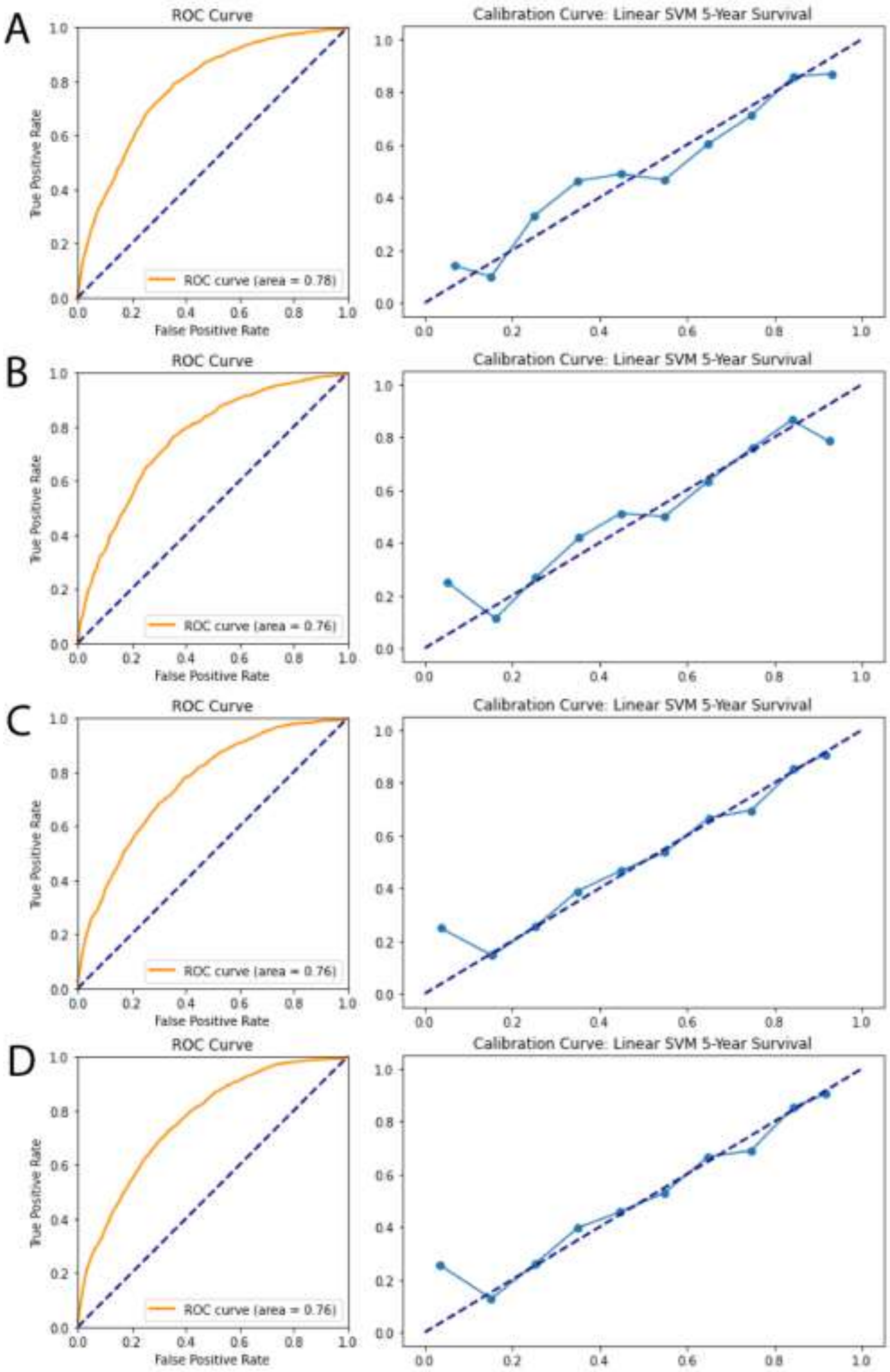

Figure 1. Receiver operating characteristic and calibration curves for 5-year survival for the internal validation of the Linear Support Vector Machine model built using the (A) *full* (B) *excluded* (C) *positive* and (D) *negative* datasets.

Table 1. Model performance on internal validation with various SEER-based datasets.

Model	c-statistic	BSL
Logistic Regression (LR)		
Full	0.79	0.207
Excluded	0.77	0.211
Positive	0.76	0.210
Negative	0.77	0.207
Linear Support Vector Machine (LVSM)		
Full	0.78	0.212
Excluded	0.76	0.215
Positive	0.76	0.212
Negative	0.76	0.211
Multi-layer Perceptron Neural Network (MLP)		
Full	0.78	0.221
Excluded	0.77	0.209
Positive	0.74	0.252
Negative	0.75	0.267

BSL; Brier score loss