Predictive Models for Metastasized Osteosarcoma Survival Outcomes in a Simulated Patient Cohort

Annie Allred

George Washington University Milken Institute School of Public Health Department of Biostatistics and Bioinformatics

Abstract

The average 5 year survival rate of metastasized osteosarcoma (OS), according to the American Cancer Society, is 5% - 30%[1]. If it can be correctly predicted whether or not a subject will survive using clinical data and information obtained from their metastasized tumor microbiome (TME), the doctor might be more willing to work with a patient, even if their odds of survival don't initially look good. A total of 11 models were run, including various forms of logistic regression, random forest, tree models, SVMs and neural networks, to predict survival outcome of 400 simulated patients[2,3]. All models had an accuracy between 73-79.25%, but random forests with classification trees and boosted classification trees with a threshold level of 0.5 had the largest accuracy. From the simulated dataset, personal and TME data held key predictors for survival.

Introduction and Problem Statement

Osteosarcoma (OS) is the most common type of bone cancer and has three stages of severity. OS tumors can be categorized into any of the three stages, but in general, localized cells are placed in stages I or II, depending on how likely they are to spread, while metastasized tumors are classified as stage III [4]. Metastasized tumors are the most dangerous and only occur when the cancer has spread to other parts of the body.

The primary aim of this study is to determine if it is possible to accurately predict whether or not a subject will die of metastatic osteosarcoma using subject data, treatment type and the OS tumor microenvironment (TME). The TME is made up of various cell types; immune cells, stromal cells, blood vessels, and extracellular matrix (ECM); and the composition varies between tumor, subject, and tumor type[5]. Furthermore, is treatment status a better predictor of death than the OS tumor microenvironment, or is an interaction of the two groupings needed?

Data Description

The dataset used in this study was simulated in part by CoPilot, although some, mainly superficial, changes were made to the generated code. The patient and clinical data was simulated using data from Song, K. *et al.*'s paper and the OS tumor microenvironment data was simulated using data from Lacinski, R. A. *et al.*'s paper[2,3]. The proportion of subjects who received surgery is the same as in Song, K. *et al.*'s paper and tumor locations, lower-extremities, upper-extremities, and spine/pelvis, were chosen from that same paper. General patient information such as age, race, and sex were randomly generated from uniform and Bernoulli distribution. Using unsupervised clustering, Lacinski, R. A. *et al.* found 45 unique cellular clusters. For each of the 45 cell clusters, a gamma distribution with shape parameter 2 was used to generate the intensities for each subject. The mean of each gamma distribution was the absolute value of the mean difference between "naïve and neoadjuvant chemotherapy exposed patients" from supplementary file 2[3].

Five of the unique cell clusters were randomly chosen to be significantly associated with survival. A risk score for each of the 400 patients is the sum of the intensities of each randomly chosen cell cluster. Probability of death increases with a patient's risk score. Age of the subject, the location of the metastasized tumor, and whether or not the subject had surgery were chosen to be significant variables. The equation: $-0.5 + 2*\text{risk\_score} + \text{age}^{1/4} - e^{\text{tumor}} - 0.75*\text{surgery}$

Methods

For this study, a total of eleven predictive classification models were created to predict the survival outcome of simulated patients with metastasized osteosarcoma. They were chosen because they seem fun to do and to compare the predictive ability of many models. Due to the small sample size, 10-fold cross validation was performed for each of the models instead of splitting the data into training and testing datasets. The `train()` function with 10-fold cross validation was used to train each of the eleven models and to get valid estimates for test error. Accuracy is the main way in which each model was assessed and the probability threshold which had the highest accuracy is reported in Table One. The sum of sensitivity and specificity is another way that the models were assessed. Interestingly, the two assessments of predictive performance agreed on the best probability threshold for three of the eleven models. Every model, except for the basic logistic regression model, had some sort of tuning grid used to find the optimal tuning parameter(s) in the `train()` function.

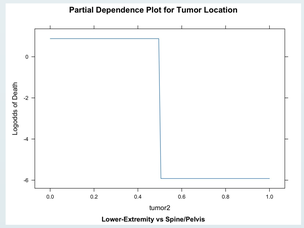
- Logistic Regression - basic model
- Logistic Regression with Ridge/LASSO penalties - handles multicollinearity and overfitting, LASSO does variable selection
- Decision Classification Tree - Highly interpretable, looks at variable importance.
- Bagging Classification Tree - reduces variance, combines weak learners, good to compare to decision tree
- Random Forest - Similar to bagging, but decorrelates trees, good to compare to bagging
- Boosted Classification Tree - combines weak learners differently than bagging, good to compare the two.
- SVMs - good when used with complex/high-dimensional data, which my dataset could be considered.
- Neural Networks - high flexibility, but due to low amount of training data, not the best in this situation.

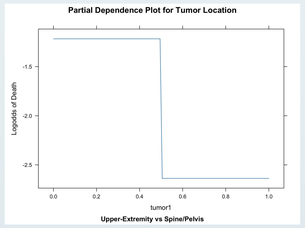
Results

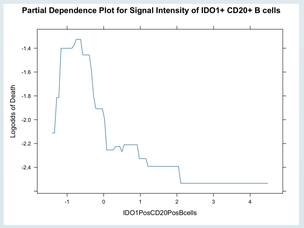
Logistic Regression:
Of the three logistic regression models, the model with a LASSO penalty had the highest predictive accuracy while the model with a ridge penalty had the lowest accuracy of the three. It is expected that a logistic regression model with a LASSO penalty would perform better than one with a ridge penalty due to the small number of predictors that have substantial coefficients. Tumor2 and tumor1 are the only predictors with coefficients greater than 1. I would have expected a plain logistic regression model to perform the worst of the three models due to the non-linear nature of the logistic equation. However, as the LSE variances are not that large in this scenario due to n > p, that might play a role in the logistic regression model with a ridge penalty performing the worst.

Neural Networks:
Of all the models, the neural network model performed the worst, but still performed well. The predictive accuracy was 3% worse than the next lowest predictive accuracy, which was the logistic regression model with a ridge penalty. The reason that the neural network model performed so poorly compared to all the other models is most likely because of the size of the training dataset. Using 10-fold CV, each training set only had 360 patients in each. Neural networks shine when the training dataset is, at minimum, in the thousands. The neural network model was one of the three models where the sum of sensitivity and specificity had the same optimal predictive threshold as accuracy.

Predictive Model	Probability Threshold	Accuracy	Sens + Spec	Parameters
Logistic Regression	0.3	0.78	1.5791	NA
Logistic Regression with Ridge Penalty*	0.3	0.76	1.6044	$\lambda = 0.0374$
Logistic Regression with LASSO Penalty	0.5	0.7825	1.5824	$\lambda = 0.0365$
Decision Classification Tree	0.4	0.7825	1.5863	$\alpha = 0.0643$
Bagging with Classification Trees	0.5	0.7825	1.5467	ntree = 500
Random Forests with Classification Trees	0.5	0.7925[A]	1.5720	ntree = 500 mtry = 44
Boosted Classification Trees	0.5	0.7925[A]	1.5489	ntree = 200 depth = 4 shrinkage = 0.1 minbucket = 5
SVM Linear Kernel*	0.4	0.79	1.6143	C = 10
SVM Polynomial Kernel	0.4	0.785	1.5967	C = 10 degree = 1 scale = 2
SVM Radial Basis Kernel	0.5	0.77	1.4648	C = 10 $\sigma = 0.005$
Neural Network*	0.5	0.73	1.4033	size = 9 decay = 0.01







Tree Models:
All four tree models, agreed on the top three most important variables, which were tumor2 (lower-extremity vs spine), tumor1 (upper-extremity vs spine), and the signal intensity of IDO1+ CD20+ B cells. As complexity of the models increase, so does accuracy. The classification decision tree is the simplest model, off of which all tree models build off of, and has lowest predictive accuracy. Bagging decision tree models are the next most complicated, and it has a slightly greater predictive accuracy. Random forest and boosted classification tree models are more complicated in different ways compared to bagging. The two models have the same, highest predictive accuracy of all models (notated with a superscript A in the table). The partial dependence plots for the three most important variables, compared to all of the predictors in the dataset, are shown above.

SVMs:
The SVM with a linear kernel had the greatest predictive accuracy of the three SVM models, followed by the polynomial kernel and then the radial basis kernel. The dataset didn't have any radial data, which may account for radial basis model low performance. However, the results for the radial basis kernel must be taken with a grain of salt as the warning "maximum number of iterations reached" occurred for every pairing in the tuning grid. The linear kernel model was one of the three models where the sum of sensitivity and specificity had the same optimal predictive threshold as accuracy. In addition, it had the third highest predictive accuracy of all the models, followed closely by the polynomial kernel model.

Conclusion

After assessing the eleven models using predictive accuracy, the top three models were random forest, boosted classification tree, and support vector machine with a linear kernel. The random forest and boosted classification tree models had the same accuracy of 79.25%. However, the random forest model performed better with respect to the sum of sensitivity and specificity compared to boosted classification tree model. In future studies, I would recommend using random forest and boosted classification tree models when using accuracy as a predictive assessor.
Based on the two recommended models, the three most important predictors, compared to all the predictors used, were tumor location, lower-extremity vs spine and upper-extremity vs spine, and the signal intensity of IDO1+ CD20+ B cells. For the tumor location predictors, the odds of living increased if the tumor was not located in the spine/pelvis. The odds of death decreased as the signal intensity of IDO1+ CD20+ B cells increased. This shows that both clinical and tumor microenvironment data are important when predicting the survival outcome of metastasized Osteosarcoma.

Citations

1. Survival Rates for Osteosarcoma. https://www.cancer.org/cancer/types/osteosarcoma/detection-diagnosis-staging/survival-rates.html.
2. Song, K. et al. Survival analysis of patients with metastatic osteosarcoma: a Surveillance, Epidemiology, and End Results population-based study. Int. Orthop. 43, 1983–1991 (2019).
3. Lacinski, R. A. et al. Spatial multiplexed immunofluorescence analysis reveals coordinated cellular networks associated with overall survival in metastatic osteosarcoma. Bone Res. 12, 55 (2024).
4. Stages and Prognostic Markers of Osteosarcoma. https://www.cancer.org/cancer/types/osteosarcoma/detection-diagnosis-staging/staging.html.
5. Anderson, N. M. & Simon, M. C. Tumor Microenvironment. Curr. Biol. CB 30, R921–R925 (2020).

I acknowledge my use of Generative AI in the preparation of this assignment in the form of CoPilot. CoPilot was used to generate data by simulating the findings from two separate papers and combining the data. I have taken all necessary steps to ensure the accuracy of the material and data I used.

YouTube Link: https://youtu.be/Ur5TJHO87Cw