Machine Learning to Predict Treatment in Oropharyngeal Squamous Cell Carcinoma

Omar Karadaghy[1], Matthew Shew[1], Jacob New[1], and Andres Bur[1]

[1]University of Kansas Medical Center

May 11, 2020

Abstract

Objective: (1) Demonstrate how machine learning can be used for prediction modeling by predicting the treatment patients with T1-2, N0-N1 Oropharyngeal Squamous Cell Carcinoma receive. (2) Assess disparities in the treatment of this population. Design: Retrospective cohort. The data was split into 80/20 distribution for training and testing, respectively. Machine learning algorithms were explored for development. Area Under the Curve, accuracy, precision, and recall were calculated for the final model. The permutation feature scores highlight significant variables within the model. Setting: National Cancer Database. Participants: Adults diagnosed with T1-2, N0-N1 Oropharyngeal Squamous Cell Carcinoma from 2004 to 2013 were eligible Main Outcome Measure: Primary treatment modality Results: Among the 19,111 patients in the study, the mean (standard deviation) age was 61.3 (10.8) years, 14,034 (73%) were male, and 17,292 (91%) were white. Surgery was the primary treatment in 9,533 (50%) cases, and radiation in 9,578 (50%) cases. The final model yielded an Area Under the Curve of 78% (95% CI, 77% to 79%), accuracy of 71%, precision of 72%, and recall of 71%. The T-stage, primary site, N-stage, grade, and type of treatment facility were impactful variables included in the model. Conclusion: Machine learning was used to predict primary treatment modality for T1-2, N0-N1 Oropharyngeal Squamous Cell Carcinoma. This study demonstrates how machine learning can be used for prediction modeling. The results also suggest treatment is influenced by clinical staging and type of treatment facility.

Key points:

- Machine learning is a novel form of analysis that is being exceedingly applied to the medical field.
- Machine learning outperforms traditional statistics in its ability to process large data input and account for variability, nonlinear interactions, and heterogeneous distributions.
- Factors influencing the decision to undergo primary surgical or primary radiation for treatment of T1-2, N0-N1 Oropharyngeal Squamous Cell Carcinoma are under investigated.
- The largest influencers were found to be tumor characteristics and type of facility that treatment was sought at.
- Applications of machine learning within Otolaryngology are expected to increase as machine learning techniques disseminate.

Introduction:

The past few decades have seen the rise of Human Papilloma Virus (HPV)-related Oropharyngeal Squamous Cell Carcinoma (OPSCC). This shift in epidemiology has led to exploration with de-escalation trials that affect our treatment of this cancer. The optimal treatment for T1-2, N0-N1 OPSCC is an ongoing debate as the current understanding of disease processes and advancing technologies are constantly changing.[1-3]

Current National Comprehensive Cancer Network (NCCN) guidelines recommend single modality treatment with either surgery or radiation for both HPV and non-HPV-related T1-2, N0-N1 OPSCC. While we await the results of several ongoing clinical trials, systematic and retrospective reviews suggest no difference in survival outcomes for the treatment of early-stage OPSCC between either modality.[1,4]

Given equivalent survival outcomes with either treatment modality in this population, little work has been conducted looking at the influence patient, socioeconomic, regional, or institutional factors have in primary treatment modality for this category of OPSCC. This question is ideally analyzed using large national data registries and a methodology equipped to analyze multiple layers of influence.

Machine learning (ML) is a novel form of analysis that uses sophisticated statistical theories to create a prediction model.[5,6] Many of the statistical principles vital to the machine learning process are similar to traditional statistical methodologies used in clinical medicine, but the primary objective of machine learning is to predict an unknown component rather that determine inferences.[5,7,8] Machine learning excels in its ability to analyze complicated interactions that exist between these variables.[56] There is growing interest in various fields of medicine to use machine learning to improve upon current methodologies.[5-7]

This study therefore seeks to utilize machine learning to create a prediction model for the primary treatment modality of patients with T1-2, N0-N1 OPSCC by examining patient, socioeconomic, regional, and institutional factors in addition to tumor factors. In doing so, this study will demonstrate how machine learning can be utilized to create prediction models in a reproducible manner, and provide insight to the variables that influence treatment patterns.

Materials and Methods:

Collection of Data

All data was collected from the National Cancer Database (NCDB), which is a joint project of the Commission on Cancer (CoC) of the American College of Surgeons (ACS) and the American Cancer Society. This nationwide hospital-based oncology data captures approximately 70% of all newly diagnosed cancers in the United States annually. Data collected is compliant with established protocols coordinated under the North American Association of Central Cancer Registries (NAACCR). All NCDB data continuously undergo integrity checks.[9]

Participants

Patients 18 years or older diagnosed with or treated for clinical T1 or T2 cancers with N0 or N1 nodal status and M0 OPSCC from 2004 to 2013 were eligible for inclusion in the study. Exclusion criteria included histologic diagnosis other than squamous cell carcinoma, advanced clinical stage disease, cases treated with palliative intent, and cases with missing information. Patients without known HPV status were also included in the study. Patients without HPV status were included in this study due to the overwhelming number (78%) of patients within the NCDB lacking HPV status.

Variables of Interest

Variables studied were divided into four separate categories: Patient, tumor, facility, and treatment characteristics. Patient characteristics include age, gender, race, and comorbid disease as calculated by the Charlson-Deyo Score[10]. Insurance status was collected and divided based on uninsured, private, or government insurance status. Tumor characteristics collected include the clinical T and N classification. The TNM editions captured in this data set include the 6th and 7th edition for which the staging of OPSCC remained the same.[11] Because the objective of the study was to predict primary treatment modality, any information derived from the time of treatment was excluded. This includes pathological TNM staging, grade, extracapsular spread, perineural invasion and other variables unless the information could clearly be identified as obtained through pre-treatment biopsy. Facility characteristics explored include whether the treating hospital is a community program, academic program, integrated network cancer program, or other. Facility location was also included to determine if geographical location affects treatment type. The NCDB divides the locations into New England, Middle Atlantic, South Atlantic, East North Central, East South Central, West North Central, West South Central, Mountain and Pacific. Primary treatment modality was divided into two categories; surgical and non-surgical treatments. Non-surgical treatments include primary radiation alone or with chemotherapy. The reason for lack of surgery, radiation, or other treatment was also collected and explored.

Machine Learning Prediction Modeling

The ML model was constructed using Azure Machine Learning Studio (Microsoft Corporation). The data was split into an 80/20 distribution, with 80% of data used to train the model, and the remaining 20% to test performance.[12,13]{Bur, 2019 #1359;Bur, 2019 #1465} Various two-class decision models were explored for development, including decision forest, decision jungle, boosted decision tree, logistic regression, and neural network. The specifications of the model were set on a parameter range in order to grant the ML the most flexibility in model development. The outcome of interest, primary treatment modality, was identified as the label.

Performance and Validation of the Prediction Model

All models were evaluated using the test data. The metrics used to assess performance include Area Under the Curve (AUC), accuracy, precision, and recall in accordance with previous recommendations for results reporting of clinical prediction models.[14] The model with the highest performance in the most performance metrics was chosen as the final model. The permutation feature importance (PFI) scores were also obtained to illuminate the most significant variables used in the model's prediction. The PFI scores are the difference in model performance determined by the AUC before and after alteration of a given dependent variable. Thus, the absolute magnitude of a PFI score reflects the impact an individual variable has on the overall performance.

Statistical Analyses

Analysis was conducted on Azure Machine Learning Studio. IBM SPSS Statistics version 25 was used to calculate the confidence intervals for the Area Under the Curve as this is not a supported function of the Azure platform. This study was declared exempt by the institutional review board of [removed for blind peer review] because no patient, provider, or hospital identifiers were examined, no protected health information was reviewed, and the analysis was retrospective.

Results:

Study Population Characteristics and Survival Information

In total, 20,830 patients were eligible for this study. The final sample size was 19,111 after excluding 1,532 patients who received treatment other than surgery or radiation, and 187 patients who received palliative treatment. The mean age of patients included in the study was 61.3 years with a standard deviation of 10.8 years. The majority of patients were male (n=14,034, 73%). Approximately 90% of the patients were white (n=17,292). Table 1 displays the full patient demographic included in this study. The primary treatment was evenly split between surgical and nonsurgical modalities. Furthermore, Table 1 displays the patient demographic of the training and testing data set.

Prediction Model Development and Validation

For model development, the classification models explored included decision forest, decision jungle, boosted decision tree, neural network, and logistic regression. Following development of several models using the above classifications, each model was applied to the test data set where performance metrics were measured. The performance scores of each model are displayed in Table 2. The decision forest classification was the most robust with an AUC of 0.78 (95% CI, 0.77 to 0.79), accuracy of 71%, precision of 72%, and recall of 71%, with the highest scoring in two out of the four measured criteria. This was followed closely by the decision jungle classification. The remaining models were noted to have declining recall performance. The ideal parameter determined by the model was minimum of 4 sample per leaf node, 128 random splits per node, maximum of 64 for depth of decision tree, and limitation to 32 different decision trees.

Prediction Performance

The PFI score is a method to evaluate the significance each individual input variable has on the overall model. The results are displayed in Table 3. The most important features are displayed in ascending order along with their corresponding importance score. In the creation of the clinical prediction model, the most important variable was patient clinical T- classification. This was closely followed by several other tumor descriptors including primary site of the cancer, clinical N- classification, grade. The results also indicate that facility type and facility ID were important variables in the creation of the model.

HPV Subanalysis

The vast majority of patients included in the study were with missing HPV status (78%). An additional analysis was conducted excluding all patients with missing HPV information to further explore the data. The machine learning models were redeveloped and retested using this new population. In all, 4,284 patients were included in the subset analysis. Full demographic information is detailed in Supplemental Table 1. Of note, primary treatment with surgical resection was slightly more common than primary treatment with radiation (60% vs. 40%). In Supplemental Table 2, the performance of the machine learning models is displayed. The decision forest again yielded the strongest model with an AUC of 75% (95% CI, 72% to 79%), accuracy of 72%, precision of 68%, and recall of 65%. The most important factor found by way of the PFI scores was patient clinical T- classification. This was closely followed by the same tumor descriptors including primary site of the cancer, clinical N- classification, and grade. The full results of the PFI analysis are displayed in Supplemental Table 3.

Discussion:

In this study, machine learning was used to create a model to predict primary treatment modality for OPSCC. Logistic regression, a more traditional statistical methodology, was employed as a reference as well. The results indicate that machine learning was able to create a robust prediction model using the variables included in the NCDB. Furthermore, the results of this study highlight that the variables most predictive of primary treatment modality are Clinical T- classification and N- classification, primary site of tumor, and type of institution where treatment is performed.

In light of the lack of sound evidence dictating optimal primary treatment modality for early-stage OPSCC, this study provides further insight about participating institutions across the nation. This study indicates that the decision to undergo primary surgery versus primary radiation is most strongly influenced by tumor characteristics, and with some influence from facility type. Our model did not find that geographical region was an important variable used to predict primary treatment modality. Previous work has demonstrated marked regional variation in pursing primary treatment with surgery for early-stage I or II cancers. This previous study found the highest surgery rates in the West North Central region, and lowest in the New England region.[15]

Machine learning is emerging in the medical literature as a novel, sophisticated methodology for predicting clinical features of interest. The advantages of ML are in its ability to process large data input and account for high levels of variability, nonlinear interactions, and heterogeneous distributions.[13,16] Use of this technology is widely used commercially, with large companies such as Netflix utilizing ML to better cater to its clientele.[17] In the medical literature, machine learning is being applied in both the clinical and basic science fields from medical imaging to genomic sequencing to predicting clinical outcomes.[13,18,19] An area where ML may be of additional value is in the analysis of large clinical data registries. Multi-institutional, national, and international data registries allow higher statistical power and open doors to address previously difficult-to-answer questions.[20] Due to the aforementioned strengths, a model developed using ML would be able to account for intricate interactions among variables.[26]

Machine learning analysis exists on a spectrum ranging from highly supervised models where all input variables and their relationships along with the desired output variable are selected by the operator to unsupervised models where the ML algorithm attempts to identify patterns of structure in unlabeled data with minimal operator input.[12] The results of this study indicate that a form of supervised learning, decision forest, yielded the strongest model. We can ascertain from this result and our previous understanding of ML that the appropriate form of data analysis is determined by the clinical question asked and the relationship between input and output variables.[21] In situations where linear problems are explored, linear or logistic regression will likely outperform ML. However, in data where nonlinear relationships and interaction terms are explored, ML will likely outperform more traditional statistics.[8,12,21]

There are several limitations that warrant discussion. To begin, this study relied on data collected from the NCDB for the development of the model. Previous studies have described the limitations using this large national registry.[22,23] Briefly, these shortcomings include incomplete patient and treatment attributes collection in the registry, and significant changes within the past decade that affect the completeness of available data. In our study, the availability of HPV data is one such example. Until recently, HPV status was not routinely collected. The HPV status for all cases in our study prior to 2010 are unknown. The decision to include those with missing HPV status was made due to NCCN guidelines for treatment of early stage-OPSCC to be primary surgery or primary radiation regardless of HPV status. However, a subanalysis was performed, which demonstrated that the same variables affect the primary treatment modality patients receive.

The final limitation for discussion is directed toward ML. In the creation of any machine learning algorithm, the process of how an algorithm determines its prediction is not available for review. This is known as the "black box" of machine learning. That is, information is input into the data and a prediction is generated, but any attempts to analyze what the impact of individual variables through an effect size or the relationship among variables is not able to be displayed in a comprehensible format.[12]

In an attempt to understand how a machine learning model produces its model, PFI scores are calculated to assess the impact of individual variables. However, interpretation of the scores is challenging. The definition of a PFI score is the absolute difference in AUC of the final model before and after altering an individual variable. Given that this is a novel metric, it is unclear what the significance of the produced value is. Furthermore, it is unknown how to compare one score to another. While two variables may have a PFI score difference of 0.01, how significant of a difference this is not understood. Therefore, this presents an additional limitation to machine learning studies as conclusions regarding individual input variables are limited to ranking. These limitations will be improved upon in the future as more investigation into prognostic patient or tumor characteristics are identified, and further work into understanding machine learning are undertaken.

Conclusion:

In this study, machine learning was used to predict the treatment an individual patient will likely undergo using pre-treatment variables obtained in the NCDB. In doing so, this study demonstrates how the use of machine learning can be broadened in prediction modeling for future analysis within clinical research. This study also highlights that the largest influencers determining the treatment for an individual patient are tumor characteristics with minor influencers including type of facility that treatment was sought at.

References

1. de Almeida JR, Byrd JK, Wu R, et al. A systematic review of transoral robotic surgery and radiotherapy for early oropharynx cancer: a systematic review. *Laryngoscope.* 2014;124(9):2096-2102.

2. Moore EJ, Olsen KD, Kasperbauer JL. Transoral robotic surgery for oropharyngeal squamous cell carcinoma: a prospective study of feasibility and functional outcomes. *Laryngoscope.*2009;119(11):2156-2164.

3. Eisbruch A, Harris J, Garden AS, et al. Multi-institutional trial of accelerated hypofractionated intensity-modulated radiation therapy for early-stage oropharyngeal cancer (RTOG 00-22). *Int J Radiat Oncol Biol Phys.* 2010;76(5):1333-1338.

4. Pedro C, Mira B, Silva P, et al. Surgery vs. primary radiotherapy in early-stage oropharyngeal cancer. *Clin Transl Radiat Oncol.*2018;9:18-22.

5. Hinton G. Deep learning—a technology with the potential to transform health care. *JAMA.* 2018.

6. Naylor C. On the prospects for a (deep) learning health care system.*JAMA.* 2018.

7. Kang J, Schwartz R, Flickinger J, Beriwal S. Machine Learning Approaches for Predicting Radiation Therapy Outcomes: A Clinician's Perspective. *Int J Radiat Oncol Biol Phys.* 2015;93(5):1127-1135.

8. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning.*Nat Methods.* 2018;15(4):233-234.

9. Bilimoria KY, Stewart AK, Winchester DP, Ko CY. The National Cancer Data Base: a powerful initiative to improve cancer care in the United States. *Ann Surg Oncol.* 2008;15(3):683-690.

10. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol.*1992;45(6):613-619.

11. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM.*Ann Surg Oncol.* 2010;17(6):1471-1474.

12. Bur AM, Shew M, New J. Artificial Intelligence for the Otolaryngologist: A State of the Art Review. *Otolaryngol Head Neck Surg.* 2019;160(4):603-611.

13. Karadaghy OA, Shew M, New J, Bur AM. Development and Assessment of a Machine Learning Model to Help Predict Survival Among Patients With Oral Squamous Cell Carcinoma. *JAMA Otolaryngol Head Neck Surg.* 2019.

14. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation.*Eur Heart J.* 2014;35(29):1925-1931.

15. Liederbach E, Lewis CM, Yao K, et al. A Contemporary Analysis of Surgical Trends in the Treatment of Squamous Cell Carcinoma of the Oropharynx from 1998 to 2012: A Report from the National Cancer Database. *Ann Surg Oncol.* 2015;22(13):4422-4431.

16. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* 2007;2:59-77.

17. Shew M, New J, Bur AM. Machine Learning to Predict Delays in Adjuvant Radiation following Surgery for Head and Neck Cancer.*Otolaryngol Head Neck Surg.* 2019:194599818823200.

18. Liu J, Wang X, Cheng Y, Zhang L. Tumor gene expression data classification via sample expansion-based deep learning.*Oncotarget.* 2017;8(65):109646-109660.

19. Pieszko K, Hiczkiewicz J, Budzianowski P, et al. Predicting Long-Term Mortality after Acute Coronary Syndrome Using Machine Learning Techniques and Hematological Markers. *Dis Markers.*2019;2019:9056402.

20. Subbarayan RS, Koester L, Villwock MR, Villwock J. Proliferation and Contributions of National Database Studies in Otolaryngology Literature Published in the United States: 2005-2016. *Ann Otol Rhinol Laryngol.* 2018;127(9):643-648.

21. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models.*J Clin Epidemiol.* 2019;110:12-22.

22. Bur AM, Holcomb A, Goodwin S, et al. Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma.*Oral Oncology.* 2019;92:20-25.

23. Boffa DJ, Rosen JE, Mallin K, et al. Using the National Cancer Database for Outcomes Research: A Review. *JAMA Oncol.*2017;3(12):1722-1728.

Hosted file

Table 1.docx available at https://authorea.com/users/319983/articles/449630-machine-learning-to-predict-treatment-in-oropharyngeal-squamous-cell-carcinoma

Hosted file

Table 2.docx available at https://authorea.com/users/319983/articles/449630-machine-learning-to-predict-treatment-in-oropharyngeal-squamous-cell-carcinoma

Hosted file

Table 3.docx available at https://authorea.com/users/319983/articles/449630-machine-learning-to-predict-treatment-in-oropharyngeal-squamous-cell-carcinoma

Hosted file

Supplemental Table 1.docx available at https://authorea.com/users/319983/articles/449630-machine-learning-to-predict-treatment-in-oropharyngeal-squamous-cell-carcinoma

Hosted file

Supplemental Table 2.docx available at https://authorea.com/users/319983/articles/449630-machine-learning-to-predict-treatment-in-oropharyngeal-squamous-cell-carcinoma

Hosted file

Supplemental Table 3.docx available at https://authorea.com/users/319983/articles/449630-machine-learning-to-predict-treatment-in-oropharyngeal-squamous-cell-carcinoma