Development of a supervised machine learning model to predict recurrence of oral tongue squamous cell carcinoma

Yasaman Fatapour, MS\textsuperscript{1}; Arash Abiri, BS\textsuperscript{1,2}; Edward C. Kuan, MD MBA\textsuperscript{2}; James P. Brody, PhD\textsuperscript{1}

\textsuperscript{1} Department of Biomedical Engineering, University of California, Irvine, Irvine, CA, USA
\textsuperscript{2} Department of Otolaryngology-Head and Neck Surgery, University of California, Irvine, Irvine, CA, USA

Corresponding author:
James P. Brody, PhD
Department of Biomedical Engineering
University of California, Irvine
3120 Natural Sciences II, Irvine, CA 92697
Phone: 949-824-2471
E-mail: jpbrody@uci.edu

Word Count: 3043
Conflicts of Interest: None declared
Abstract

Objective: Despite diagnostic advancements, development of reliable prognostic systems for assessing risk of cancer recurrence still remains a challenge. In this study, we developed a novel framework to leverage the expansive Surveillance, Epidemiology, and End Results (SEER) database to generate highly representative machine learning prediction models for oral tongue squamous cell carcinoma (OTSCC) cancer recurrence.

Materials and Methods: Using our framework, we identified cases of 5- and 10-year OTSCC recurrence from the SEER database. Cases were split into training (80%) and test (20%) sets for model development and testing. Four classification models were trained by using the H2O artificial intelligence platform, whose performances were assessed according to their accuracy, recall, precision, and the area under the curve (AUC) of their receiver operating characteristic (ROC) curves. By evaluating Shapley additive explanations contribution plots, feature importance was studied.

Results: Of 130,979 patients, 36,042 (27.5%) were female and the mean (SD) age was 58.2 (13.7) years. The Gradient Boosting Machine model performed the best, achieving 81.8% accuracy, 0.75 AUC, 83.0% recall, and 97.7% precision for 5-year prediction. Moreover, 10-year predictions demonstrated 80.0% accuracy, and 94.0% precision. The number of prior tumors, patient age, site of cancer recurrence, and tumor histology were the most significant predictors.

Conclusion: Implementation of our novel SEER framework enabled successful identification of patients with OTSCC recurrence, with which highly accurate and sensitive prediction models were generated. Thus, we demonstrated our framework’s potential to be applied to various cancers for building generalizable screening tools to predict tumor recurrence.
**Keywords**: oral tongue squamous cell carcinoma; cancer recurrence; machine learning; oral cancer
Total number of pages

1. Text pages: 25 pages
2. Tables: 2 tables
3. Figures: 3 figures
Introduction

Oral tongue squamous cell carcinoma (OTSCC) is a common head and neck neoplasm that accounts for approximately 1% of new cancer cases diagnosed in the United States each year. Despite advancements in cancer therapeutics and surgical techniques, the worldwide incidence of OTSCC is on the rise and adequate OTSCC management still remains a challenge, with patient 5-year survival rates averaging at about 50%. With recent studies reporting recurrence rates as high as 32.7%, further investigations aimed at optimizing treatment regimens and post-therapy follow-up will be critical to enhancing patient outcomes.

The advent of machine learning (ML) and its adoption by the medical community has enabled unique perspectives and solutions to numerous medical challenges. Over the past decade, scientific efforts have demonstrated the utility of machine learning in guiding cancer diagnosis and management in a variety of medical fields, including general surgery, neurosurgery, and otolaryngology. Specifically, many studies have applied machine learning techniques for predicting tumor diagnosis, tumor recurrence and patient survival in the context of various cancers. Recently, Alabi et al. and Karadaghy et al. demonstrated the capacity for ML to elucidate models for predicting recurrence and survival, respectively, in OTSCC patients. However, as with many of their predecessors, these studies were limited by the small samples of patients from which their models were trained.

Over the past two decades, the widespread shift towards the use of electronic medical records has resulted in a rapid accumulation of digital medical data, from which large administrative registries have been formed. The Surveillance, Epidemiology, and End Results (SEER) program, in particular, provides one of the largest cancer databases in the United States and represents nearly 48% of the national population. Recently, ML experts have been able to
leverage the expansive nature of the SEER database to generate more precise and representative models for predicting patient survival. However, to date, there are a paucity of studies that have attempted to utilize this database for predicting the recurrence of a cancer following treatment and complete remission.

Thus, in this study, we developed a novel algorithm to identify cases of cancer recurrence in the SEER database, from which we generated ML models to accurately predict 5- and 10-year locoregional OTSCC recurrence. By using simple and commonly acquired prognostic markers as the basis of our models, we enabled our system to be more accessible and easily adoptable by a wide range of practitioners. Furthermore, we leveraged our nationally representative ML models to accurately classify patients into low- and high-risk categories. Hence, our system not only lays a foundation for future ML efforts for predicting cancer recurrence using the SEER database, but also serves as an accurate data-driven tool for predicting OTSCC recurrence, with implications in guiding cancer management and follow-up medical care.
Methods

A novel strategy was implemented for extracting cases from the SEER database with the goal of identifying locoregional recurrence of cancer within 5-year and 10-year periods. As further detailed in the sections below, SEER*Stat version 8.3.9 (Surveillance Research Program, National Cancer Institute) was used to extract data from 18 SEER registries from 2000 to 2018. After processing the data and extracting variables of interest, dataset grouping, feature extraction, and training and validation of the model was performed (Figure 1).

A. SEER DATABASE QUERY (DATA SOURCE)

The 2000-2018 SEER database is a deidentified registry that reports cancer incidence and survival data on approximately 48% of the national population, serving as one of the largest and most comprehensive efforts for tracking oncological cases within the U.S. Due to the massive scale of available data, this work utilized SEER as its target database. Due to the anonymized and public nature of the SEER database, this study was exempt from University of California Irvine Institutional Review Board approval.

The database was queried for patients diagnosed with OTSCC using the International Classification of Disease for Oncology, 3rd Edition (ICD-O-3) topography codes for the oral tongue (C02.0-C02.9) and histology/behavior codes for squamous cell carcinoma (SCC; 8010/3, 8020/3, 8021/3, 8070/3, 8071/3, 8072/3, 8073/3, 8074/3, 8082/3). The following demographic and clinical variables of interest were used for training our machine learning models: age, sex, race, marital status, year of diagnosis, number of prior tumors, tumor site (e.g., ventral surface of tongue, dorsal surface of tongue, border of tongue), histology, tumor grade, T/N/M stage, and administered treatments (i.e., surgery, radiation, chemotherapy). To account for variant-specific
OTSCC behavior, histology was stratified into the following prognostic categories: nonkeratinizing SCC with maturation, undifferentiated nonkeratinizing SCC, differentiated nonkeratinizing SCC, and keratinizing SCC.\textsuperscript{25} Furthermore, each case contained a sequence number that provided information on the number of all reportable primary tumors that had occurred over the lifetime of a patient. This variable was used to calculate the “Number of prior tumors”, which was defined as the sequence number minus one. All cases with unknown or missing sociodemographic or outcome variables were excluded.

**B. Patient Grouping and Feature Extraction**

Each individual case in the SEER dataset was defined by a unique patient identification number. Cases were first grouped according to their patient IDs, and subsequently sorted within their groups using their sequence numbers. Next, a series of validations was performed for all patients and their respective cases. These validations focused on minimizing errors in later classification steps by eliminating conditions where the “state of recurrence” (recurrence = true/false) could not be determined with absolute confidence based on the available SEER data. The following validations were implemented:

1) The oral tongue should be the primary site of the first case for each patient.

2) All cases corresponding to patients with missing or unknown value for any variable critical for analysis, including Total Number of Malignant Tumors, Sequence Number, Survival Months, and Year of Diagnosis, were filtered out.

In the final step of the algorithm, we computed the target outcome variable, “Will Recure”. This variable, which was computed for each individual case, defined whether or not a case would recur in locoregional sites within the defined period of time (5 and 10 years). Of
note, due to SEER coding guidelines, a recurrence that occurred at the exact same topographical code as its prior incident case, was not reported in the database and, thus, unavailable for analysis. A non-recurrence was defined as a patient that had only one primary tumor and survived longer than the target window (e.g., 5 years). Conversely, if there was another recurrence of the cancer within the target window and in the same region as the initial tumor, then the case was marked as “Will Recur” = true. It is worth noting that, based on the algorithm above, the last case for a patient with multiple primary tumors (i.e., multiple case) would be marked as will not recur if the patient survived longer than the target window without another recurrence of cancer. This is critical as it tends to indicate a successful treatment.

C. ML TRAINING & VALIDATION (BALANCING, UNDER SAMPLING, NUMBER OF RUNS AND DISTRIBUTION OF DATA)

We used the H2O AI platform (H2O.ai, Inc, Mountain View, CA) in conjunction with R statistical computing environment (version 3.6.1; The R Foundation for Statistical Computing) to train and test numerous machine learning models with the goal of identifying the best model for prediction of locoregional recurrence of OTSCC. In order to properly validate and test each model, the dataset was split into training (80%) and test (20%) sets. H2O’s Automl function was used to run through different machine learning algorithms and evaluate various hyperparameters for each algorithm. 26

Due to the unbalanced nature of the dataset (i.e., much fewer recurrence cases compared to non-recurrence), two approaches for balancing data were evaluated. The first was oversampling, which involved synthesizing new examples from the existing samples for the minority class. 23 The down side of oversampling is that it introduces risk of overfitting and/or
introducing mathematically valid, yet logically non-sensical sample sets. The second approach was under-sampling, which involved randomly selecting examples from the majority class to remove from the training dataset. In general, under sampling is the preferred method, particularly for large dataset.\textsuperscript{27,28} In this case, application of the massive SEER data set helped make utilization of the under-sampling approach a reality, further strengthening the accuracy of the final model. H2O was executed with a 5-fold cross validation and configured for a maximum runtime of 600 seconds. For each ML model, 5 different runs were executed, and the average performances of the top four ML models were compared using the areas under the curves (AUCs) of the receiver operating characteristic (ROC) curves.
Results

A. Study Population Characteristics and Cancer Recurrence Information

A total of 136,826 cases were extracted from SEER dataset which represented 130,979 unique patients. Two models were trained, one focusing on locoregional recurrence of OTSCC in a 5-year period and the other, a 10-year period. In the 5-year analysis, 14,530 patients met the inclusion criteria, of which 657 suffered from a locoregional recurrence. For the 10-year analysis, 7,100 patients met the inclusion criteria, of which 971 experienced a locoregional recurrence. It is worth noting that only patients alive within the follow-up period (5- or 10-years) were considered in our analyses. Table 1 shows a summary of predictors that were used for training the machine learning model.

B. Model Prediction and Development (Performance Metric for the Algorithm)

To identify the most predictive model, the AUC of the ROC curve was used as a metric to compare the performance of four machine learning algorithms, Generalized Linear Model (GLM), Gradient Boosting Machine (GBM), Distributed Random Forest (DRF), and deep learning (artificial neural network, Figure 2).

The performance metrics of top four ML model are shown and compared in Table 2. GBM classification model with AUC of 0.75 (0.01) and 0.74 (0.02) outperformed all other models for both 5-year prediction and 10-year prediction respectively. Of note, the accuracy, recall, and precision of the model can be calculated at different thresholds within the graph of the ROC curve. Thus, the optimum threshold for each model can vary depending on the definition and application of the classification problem. For example, a screening tool may require high recall and precision. For this proof-of-concept effort, we focused on using the model as a
screening tool and, therefore, aimed to increase recall without major sacrifice of accuracy. Therefore, the best overall performance for predicting OTSCC recurrence was achieved by the GBM model with 81.8% accuracy, 83.0% recall, and 97.7% precision for 5-year prediction, and 80.0% accuracy, 82.8% recall, and 94.0% accuracy for 10-year prediction.

In addition to performance metrics of the model, we were also interested on the impact of each individual feature on the predictive outcome. The Shapley Additive exPlanations contribution plot (SHAP) illustrates how the GBM model arrived at its results (Figure 3) and explored the non-linearity effects of features on model. It ranked (from top to bottom) the importance of each feature in a predictive model based on all the possible pairs of coalitions between predictors of the model. A higher importance score was indicative of a higher contribution to the model’s predictive ability. As shown, the number of prior tumors, age, and tumor site were the most important factors in determining the probability of locoregional recurrence of OTSCC.
Discussion

In this study, we developed a novel framework for identifying cases of cancer recurrence from the SEER database with which a generalizable and highly representative machine learning model could be generated. We demonstrated the utility of this framework by developing ML models that predicted 5- and 10-year cancer recurrence with high accuracy and precision using a large population-based cohort of OTSCC patients. Specifically, of the four ML algorithms that we employed, the GBM-based model showed the most promise, demonstrating accuracies of 82% and 80% for 5-year and 10-year recurrence, respectively. Of note, we observed a recurrence rate of ~5%, which was lower than the 16-33% recurrence rate that has been previously reported. This was due to the stringent exclusion criteria that we applied, which required that patients with certain missing or unknown case information be excluded from analysis. However, we do not anticipate this lower prevalence to have influenced our findings since, unlike traditional regression techniques that compute likelihood or risk scores based on a sample’s observed event rate, our machine learning model was trained using an under-sampling approach on the majority class (non-recurrence) in order to be tolerant of deviations from the true population prevalence rates. Ultimately, by using simple and widely accessible demographic and clinical variables as the basis for model training, our sensitive prediction model shows promise in serving as a screening tool to assist clinicians in managing OTSCC patients during and after their treatment course.

Although significant progress has been made in cancer diagnostics and treatments, the prognosis of OTSCC is still poor, with many patients experiencing cancer recurrence and surviving less than 10 years after their initial diagnoses. By developing a predictive screening tool, treatment teams can be better informed of a patient’s risk for cancer recurrence and modify...
their management strategy accordingly. Additionally, the mortality rate in recurrent cases of OTSCC is highly dependent on the time of diagnosis, with early detection of recurrence being associated with reduced mortality.\textsuperscript{33,34} By using our highly representative and sensitive classification models, clinicians can be better informed of which patients are at higher risk for OTSCC recurrence and cater their management and follow-up to ensure timely diagnosis if a recurrence were to occur.

In our analysis, we used SHAP to explain the predictions made by the Gradient Boosting model and interpret the tangled nonlinear relationships between features and local regional recurrence of OTSCC. Consequently, we found that the number of prior tumors, patient age, tumor site, chemotherapy, tumor histology and tumor grade were consistently the most influential features in predicting cancer recurrence. Thus, by developing an artificial intelligence (AI) model in the context of a highly representative population for cancer recurrence and analyzing the nonlinear effect of features by the SHAP method, we found some of the features to be more prognostic compared to those that have been traditionally considered major prognostic factors in oral tongue cancer recurrence, such as lymphatic invasion or T-stage.\textsuperscript{8,31} Importantly, these findings do not discount the prognostic importance of previously reported clinical factors, but rather highlights certain factors that may be generally considered highly prognostic across a more diverse and heterogeneous patient population.

In a recent institutional study, Alabi et al. similarly demonstrated success in predicting locoregional recurrence in OTSCC. However, despite their impressive results, their models were trained using only 217 cases of early-stage OTSCC, which largely limited their system’s applicability to more advanced tumors and its external validity against the general population, where the spectrum of disease behavior and progression is much more diverse than what is
experienced at a single institution. Interestingly, the authors found that certain specialized histopathological parameters, such as lymphocyte host response, pattern of invasion, depth of invasion, and perineural invasion, were particularly important features in their prediction models. Owing to the limitations of the SEER database, our models were trained without using these clinical features. While the lack of dependence on these specialized histopathological parameters expanded the accessibility of our system to a broader range of clinical facilities where such information may not be readily available, consideration of these features may be warranted in future generations of ML models where higher prediction accuracy in lieu of increased accessibility is desired.

Previous studies have reported on the significance of genetic predisposition in head and neck squamous cell carcinoma (HNSCC). Moreover, genetic and environmental factors, including a history of prior head and neck cancer, have been shown to be associated with recurrence of HNSCC. The influence of patient age on prognosis has also been previously established. In a large retrospective study of OTSCC patients, Mukdad et al. demonstrated that older patients were associated with more advanced disease and worse survival. It was hypothesized that this worse prognosis was partly due to a tendency for clinicians to more aggressively treated younger patients with multimodality therapy. Interestingly, younger patients were also observed to less frequently present with metastatic lymph nodes. Indeed, survival and recurrence rates have been reported to be largely influenced by the presence of nodal disease. As such, cancer recurrence at a regional site can be suggestive of more aggressive disease with tendency to recur following treatment. In a cohort study, Wolfer et al. suggested that aggressive neoplastic behavior is strongly dictated by tumor histology. Specifically, the degree of keratinization in oral squamous cell carcinoma was demonstrated to be an important
prognostic factor for recurrence and survival. Other recent studies have reached similar conclusions and have even created recurrence risk models on the sole basis of histological parameters.\textsuperscript{41–44}

To our knowledge, this is one of the first studies to develop an algorithm to identify cases of cancer recurrence from the expansive and widely used SEER database, laying a basis for future investigations across a variety of medical fields. Through the use of this novel framework, we also presented one of the first machine learning-based classification models that accurately predicted 5- and 10-year recurrence in OTSCC patients using only commonly available demographic and clinical features. There are, however, limitations to this study worth mentioning. Since patients were extracted from a de-identified national database, the data may be susceptible to information bias. Additionally, despite including a number of sociodemographic and clinical variables in our models, certain histopathological (e.g., lymphocyte host response, perineural invasion, depth of invasion) and clinical features (e.g., timing of treatments, radiation dose, HPV status) could not be accounted for due to limitations in the data available in the SEER database.
Conclusions

In this study, we developed a novel framework for identifying cases of cancer recurrence from the SEER database. Using a population-based sample of over 130,979 patients, we developed several highly accurate and sensitive machine learning models to predict OTSCC recurrence. Despite the use of simple and commonly available prognostic markers as the sole features for model training, our GBM-based model was nonetheless able to achieve prediction accuracies of 82% and 80% for 5- and 10-year cancer recurrence, respectively. With our framework’s ability to be applied to a wide variety of cancers, we believe that this tool can have significant implications in future oncologic research efforts aimed towards improving disease management and optimizing patient outcomes.
Acknowledgment

This research was partially supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number T32GM008620. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
References


doi:10.5732/cjc.012.10219


doi:10.1109/EMBC46164.2021.9630556


44. El-Mofty SK. Histopathologic risk factors in oral and oropharyngeal squamous cell
## Tables

Table 1: Summary of the sociodemographic and clinical predictors used in developing the ML models for predicting OTSCC recurrence.

<table>
<thead>
<tr>
<th>Variable</th>
<th>5-Year (N = 14995)</th>
<th>10-Year (N=7342)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. (%)</td>
<td>No. (%)</td>
</tr>
<tr>
<td>Mean Age, yrs. (SD)</td>
<td>58.4 (11.5)</td>
<td>56.2 (11.5)</td>
</tr>
<tr>
<td>Sex</td>
<td>Male 10,636 (72.0)</td>
<td>4,075 (67.7)</td>
</tr>
<tr>
<td></td>
<td>Female 4,129 (28.0)</td>
<td>1,943 (32.3)</td>
</tr>
<tr>
<td>Race</td>
<td>White 13261 (89.8)</td>
<td>5991 (90.1)</td>
</tr>
<tr>
<td></td>
<td>Black 706 (4.8)</td>
<td>270 (4.1)</td>
</tr>
<tr>
<td></td>
<td>Asian 798 (5.4)</td>
<td>387 (5.8)</td>
</tr>
<tr>
<td>Marital Status</td>
<td>Single 5056 (34.2)</td>
<td>2040 (30.7)</td>
</tr>
<tr>
<td></td>
<td>Married 9709 (65.8)</td>
<td>4608 (69.3)</td>
</tr>
<tr>
<td>Number of Prior Tumors</td>
<td>0 14051 (95.2)</td>
<td>6324 (95.1)</td>
</tr>
<tr>
<td></td>
<td>1 496 (3.4)</td>
<td>232 (3.5)</td>
</tr>
<tr>
<td></td>
<td>2 161 (1.1)</td>
<td>77 (1.2)</td>
</tr>
<tr>
<td></td>
<td>3 46 (0.3)</td>
<td>14 (0.2)</td>
</tr>
<tr>
<td></td>
<td>4+ 11 (0.1)</td>
<td>1 (0.0)</td>
</tr>
<tr>
<td>Histology</td>
<td>Nonkeratinizing SCC with maturation 11468 (77.7)</td>
<td>5276 (79.4)</td>
</tr>
<tr>
<td></td>
<td>Undifferentiated nonkeratinizing SCC 86 (0.6)</td>
<td>39 (1.0)</td>
</tr>
<tr>
<td></td>
<td>Differentiated nonkeratinizing SCC 824 (5.6)</td>
<td>288 (4.3)</td>
</tr>
<tr>
<td></td>
<td>Keratinizing SCC 2286 (15.5)</td>
<td>993 (15.0)</td>
</tr>
<tr>
<td></td>
<td>SCC NOS 101 (0.7)</td>
<td>52 (1.0)</td>
</tr>
<tr>
<td>Tumor Grade</td>
<td>Well-differentiated 2262 (18.8)</td>
<td>1067 (19.7)</td>
</tr>
<tr>
<td></td>
<td>Moderately differentiated 5752 (47.8)</td>
<td>2585 (47.6)</td>
</tr>
<tr>
<td></td>
<td>Poorly differentiated 3896 (32.4)</td>
<td>1710 (31.5)</td>
</tr>
<tr>
<td></td>
<td>Undifferentiated 117 (1.0)</td>
<td>64 (1.2)</td>
</tr>
<tr>
<td>T-Stage</td>
<td>T1 4443 (46.7)</td>
<td>1594 (50.0)</td>
</tr>
<tr>
<td></td>
<td>T2 3274 (34.4)</td>
<td>1109 (34.8)</td>
</tr>
<tr>
<td></td>
<td>T3 1013 (10.6)</td>
<td>262 (8.2)</td>
</tr>
<tr>
<td></td>
<td>T4 784 (8.2)</td>
<td>221 (6.9)</td>
</tr>
<tr>
<td>N-Stage</td>
<td>N0 5110 (45.5)</td>
<td>1918 (48.3)</td>
</tr>
<tr>
<td></td>
<td>N1 1968 (17.5)</td>
<td>764 (19.2)</td>
</tr>
<tr>
<td></td>
<td>N2 3847 (34.3)</td>
<td>1187 (29.9)</td>
</tr>
<tr>
<td></td>
<td>N3 296 (2.6)</td>
<td>102 (2.6)</td>
</tr>
<tr>
<td>M-Stage</td>
<td>M0 11200 (99.3)</td>
<td>3913 (99.2)</td>
</tr>
<tr>
<td></td>
<td>M1 75 (0.7)</td>
<td>30 (0.8)</td>
</tr>
<tr>
<td>Surgery</td>
<td>Yes 6125 (41.8)</td>
<td>2506 (37.5)</td>
</tr>
<tr>
<td></td>
<td>No 8519 (58.2)</td>
<td>4185 (62.5)</td>
</tr>
<tr>
<td>Radiation</td>
<td>Yes 8965 (60.7)</td>
<td>3811 (57.3)</td>
</tr>
<tr>
<td></td>
<td>No 5800 (39.3)</td>
<td>2837 (42.7)</td>
</tr>
<tr>
<td>Chemotherapy</td>
<td>Yes 6598 (44.7)</td>
<td>2632 (39.6)</td>
</tr>
<tr>
<td></td>
<td>No 8167 (55.3)</td>
<td>4016 (60.4)</td>
</tr>
<tr>
<td>SCC: Squamous Cell Carcinoma; NOS: Not Otherwise Specified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Values are based on the number of cases.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Performance metrics of the top 4 machine learning models for predicting 5- and 10-year cancer recurrence. The GBM model exhibited the highest AUC and accuracy for both prediction windows.

<table>
<thead>
<tr>
<th>Prediction Window</th>
<th>Classification Model</th>
<th>AUC (SD)</th>
<th>Accuracy % (95% CI)</th>
<th>Recall % (SD)</th>
<th>Precision % (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5 Years</strong></td>
<td>GBM</td>
<td>0.75 (0.01)</td>
<td>81.8 (79.7-83.9)</td>
<td>83.0 (0.02)</td>
<td>97.7 (0.002)</td>
</tr>
<tr>
<td></td>
<td>GLM</td>
<td>0.73 (0.02)</td>
<td>77.4 (74.5-80.2)</td>
<td>78.1 (0.03)</td>
<td>98.0 (0.002)</td>
</tr>
<tr>
<td></td>
<td>DRF</td>
<td>0.72 (0.03)</td>
<td>72.8 (69.8-75.7)</td>
<td>73.3 (0.02)</td>
<td>97.8 (0.003)</td>
</tr>
<tr>
<td></td>
<td>Deep Learning</td>
<td>0.70 (0.04)</td>
<td>82.1 (74.7-89.6)</td>
<td>83.5 (0.06)</td>
<td>97.6 (0.002)</td>
</tr>
<tr>
<td><strong>10 Years</strong></td>
<td>GBM</td>
<td>0.74 (0.02)</td>
<td>80.0 [75.3, 84.1]</td>
<td>82.8 (0.04)</td>
<td>94.0 (0.004)</td>
</tr>
<tr>
<td></td>
<td>GLM</td>
<td>0.73 (0.02)</td>
<td>78.4 [74.2, 82.7]</td>
<td>81.0 (0.04)</td>
<td>94.3 (0.002)</td>
</tr>
<tr>
<td></td>
<td>Deep Learning</td>
<td>0.71 (0.02)</td>
<td>74.4 [70.1, 78.8]</td>
<td>76.6 (0.04)</td>
<td>94.0 (0.002)</td>
</tr>
<tr>
<td></td>
<td>DRF</td>
<td>0.69 (0.01)</td>
<td>70.6 [68.0, 73.3]</td>
<td>72.2 (0.02)</td>
<td>93.8 (0.004)</td>
</tr>
</tbody>
</table>

AUC: Area Under Curve; GBM: Gradient Boosting Machine; GLM: Generalized Linear Model; DRF: Distributed Random Forest

Performance metrics were reported as the average of 5 runs.
Figures Legends

**Figure 1: Schematic of data processing and model development.** The model development process, data cleaning, and machine learning steps in R studio and H2O.ai tool.

**Figure 2: ROC plots of four developed ML models.** Performance of Gradient Boosting Machine (GBM), Generalized Linear Model (GLM), Distributed Random Forest (DRF), and deep learning (artificial neural network) models in predicting (a) 5-year and (b) 10-year OTSCC recurrence. Patient’s data were split into 80% training set and 20% test set and 5-fold cross validation were done in each run.

**Figure 3: The Shapley Additive exPlanations contribution plots (SHAP) for the GBM model.** SHAP plots of (a) 5-year and (b) 10-year prediction models. All pairs of coalitions between features of ML model were calculated and feature’s importance were ranked from top to bottom.