

**PREDICTING ORAL CANCER-RELATED MORTALITY AMONG  
ADULTS IN THE UNITED STATES**

---

A Dissertation  
Submitted to  
the Temple University Graduate Board

---

In Partial Fulfillment  
of the Requirements for the Degree  
MASTER OF SCIENCE

---

by  
Aavishi Arora  
Diploma Date (August 2024)

Examining Committee Members:

Dr. Chukwuebuka Ogwo, Thesis Advisor, Oral Health Sciences

Dr. Sungwoo Lim, Oral Health Sciences

Dr. Pooja Gangwani, Oral and Maxillofacial Surgery

## ABSTRACT

**Objectives:** To predict oral cancer-related mortality among adults in the United States and identify the predictors of oral cancer-related mortality using the Machine Learning Approach.

**Methods:** We extracted data for 8,176 participants from the SEER database (1975 to 2022). A series of 38 demographic, clinicopathological, and lifestyle factors were extracted from the SEER database along with the outcome variable Oral Cancer-Related Mortality (OCRM) coded as “Died from Oral Cancer” and “Alive/Died from Other Causes.” The data were pre-processed using recipe packages in R. Machine Learning (ML) models-extreme gradient boosting (XGBOOST), Lasso Regression, and K-nearest neighbor were used to perform prediction of oral cancer prognosis under five-fold cross-validation to prevent overfitting or underfitting of the data. Model performance was evaluated using the Brier score, area under the curve (AUC), specificity, sensitivity, and accuracy. ML model was performed using MachineShop Package in R.

**Results:** The study participants were 63% male and predominantly non-Hispanic white (71%). 7444 participants were alive or dead of other causes and 732 were dead due to cancer. Across all models, XGBoost ML model performed the best with a Brier Score of 0.0677, an accuracy of 91%, a 13% kappa statistic, an ROC AUC of 84%, a sensitivity of 99%, and less than 1% specificity. Out of 38 variables assessed, 17 were found to be the most important predictors of OCRM. The most important predictors of OCRM (in descending order) were cancer stage group, age, T stage, Lymph node surgery, cancer site, tumor rarity, N stage, marital status, radiation, income, grade, lymph node size,

surgery radiation sequence, race, histology, the sequence number of multiple primary cancers, side of a paired organ which tumor originated from.

**Conclusion:** Our Machine-Learning model was effective in predicting oral cancer mortality using clinicopathological variables from the national cancer registry.

**Keywords:** Machine learning, metastasis, oral cancer, prediction, squamous cell carcinoma, SEER database.

## **DEDICATION**

This thesis is dedicated to my mentors, family, and friends. Your longstanding belief in me has been the driving force of my success. The wisdom and guidance of my mentors have been instrumental in my professional and personal growth.

## **ACKNOWLEDGMENTS**

I would like to thank my main committee members Dr. Chukwuebuka Ogwo, Dr. Sungwoo Lim and Dr. Pooja Gangwani for helping me throughout the process. Thank you for your guidance, dedication, and time. I would also like to thank my family and close friends for their steady support and encouragement throughout the course of my project.

# Table of Contents

<b>ABSTRACT.....</b>	<b>II</b>
<b>DEDICATION.....</b>	<b>IV</b>
<b>ACKNOWLEDGMENTS .....</b>	<b>V</b>
<b>CHAPTER 1.....</b>	<b>10</b>
<b>INTRODUCTION.....</b>	<b>10</b>
<b>CHAPTER 2.....</b>	<b>14</b>
<b>LITERATURE REVIEW .....</b>	<b>14</b>
2.1 Introduction.....	14
2.2 Integration of the individual articles.....	23
2.3 Discussion of reviewed articles: .....	26
2.4 Conclusion .....	28
2.5 Gaps in the studies reviewed: .....	29
<b>CHAPTER 3.....</b>	<b>31</b>
<b>MATERIAL AND METHODS .....</b>	<b>31</b>
3.1 Data Source.....	31
3.2 Prognostic Features.....	32
3.3 Detailed Definition of Independent Variables .....	32
3.4 TMN classification.....	45
Clinical Significance of TMN classification.....	46
3.5 Inclusion And Exclusion Criteria.....	46
3.6 Statistical Analysis.....	47
3.6.a Data Cleaning.....	47
3.6.b Descriptive Analysis.....	48
3.7 Machine Learning (ML) .....	48
3.7.a Definition of Machine Learning Models.....	48
3.7.b Data pre-processing.....	49
3.7.c Prediction modeling .....	50
<b>CHAPTER 4.....</b>	<b>51</b>
<b>RESULTS .....</b>	<b>51</b>
<i>Table 1. Sociodemographic Features .....</i>	<i>51</i>
<i>Table 2. Clinicopathological &amp; Histological Features .....</i>	<i>53</i>
<i>Table 3. Treatment Options of Head &amp; Neck Cancer Patients .....</i>	<i>58</i>
<i>Table 4. Association between Oral Cancer-Related Mortality and Predictors of Oral Cancer mortality.....</i>	<i>60</i>

<b>PREDICTIVE MODELLING FOR ORAL CANCER RELATED MORTALITY.....</b>	<b>66</b>
<i>Table 5. Variable Importance Table - Xgboost .....</i>	<i>68</i>
<i>Table 6. Variable Importance Table - Lasso Regression .....</i>	<i>71</i>
<i>Table 7. Variable Importance Table - Random Forest .....</i>	<i>73</i>
<i>Table 8. Variable Importance Table - K- Nearest Neighbors .....</i>	<i>75</i>
<i>Table 9. Comparison Of Machine Learning Models - extreme gradient boosting machine versus lasso regression versus random forest versus K-Nearest Neighbour .....</i>	<i>76</i>
<b>CHAPTER 5 .....</b>	<b>78</b>
<b>DISCUSSION .....</b>	<b>78</b>
5.1 Strength and Weakness .....	81
5.2 Future Directions .....	82
5.2.a <i>Development of a Personalized Diagnosis App for Dental Health: .....</i>	<i>82</i>
5.2.b <i>Integration of Genomic Data for Enhanced Cancer Analysis and Risk Assessment: .....</i>	<i>83</i>
5.2.c <i>Implementation of Large Language Model (LLM) for Clinician Decision Support: .....</i>	<i>83</i>
<b>CHAPTER 6 .....</b>	<b>85</b>
<b>CONCLUSION .....</b>	<b>85</b>
<b>REFERENCES .....</b>	<b>86</b>

## LIST OF TABLES

Table	Page
1. Table 1. Sociodemographic Features .....	51
2. Table 2. Clinicopathological & Histological Features.....	53
3. Table 3. Treatment Options of Head & Neck Cancer Patients.....	58
4. Table 4. Association between Oral Cancer-Related Mortality and Predictors of Oral Cancer mortality .....	60
5. Table 5. Variable Importance Table - Xgboost .....	68
6. Table 6. Variable Importance Table - Lasso Regression.....	71
7. Table 7. Variable Importance Table - Random Forest.....	73
8. Table 8. Variable Importance Table - K- Nearest Neighbors .....	75
9. Table 9. Comparison Of Machine Learning Models - extreme gradient boosting machine versus lasso regression versus random forest versus K-Nearest Neighbour .....	76

## LIST OF FIGURES

Figure	Page
1. Figure 1. Variable Importance Plot - Xgboost .....	67
2. Figure 2. Calibration Plot - Xgboost.....	68
3. Figure 3. Variable Importance Plot - Lasso Regression .....	70
4. Figure 4. Calibration Plot - Lasso Regression.....	70
5. Figure 5. Variable Importance Plot - Random Forest.....	72
6. Figure 6. Calibration Plot - Random Forest .....	72
7. Figure 7. Variable Importance Plot- K- Nearest Neighbors .....	74
8. Figure 8. Calibration Plot- K- Nearest Neighbors.....	75
9. Figure 9. Boxplot representation of performance comparison and selection of Machine Learning Models.....	76

## CHAPTER 1

### INTRODUCTION

Oral cancer is defined as cancer that develops in the oral cavity which is the anatomical space that lies between an imaginary coronal plane drawn from the junction of the soft and hard palate and the circumvallate papillae of the tongue to the vermillion of the lips. Common sites in which oral cancer lesions occur include the lip, tongue, floor of the mouth, gingiva, alveolar mucosa, palate, buccal/labial mucosa, maxilla, and mandible. Some of the most common types of oral cancer are odontogenic tumors, epithelial tumors, hematologic tumors, bone tumors, salivary gland tumors, and mesenchymal tumors (Wong, 2018).

In the United States, head and neck cancer accounts for 2.8% of all new cancer cases. There are anticipated to be 11,580 cancer-related deaths in 2023, along with an expected 54,540 new instances of the oral cavity and pharynx cancer according to the American Cancer Society (National Cancer, 2023). Oral cancer will affect 11.5 adults out of every 100,000 people (SEER, 2023). In 2019, about 279,000 men and 131,000 women had oral cancer and it was noted that patients suffering from oral cancer were in the old age groups i.e. 50 years or older. (NCI, 2023) An estimated 424,284 people were living with oral cavity and pharynx cancer in the United States in 2020 (National Cancer, 2023). The incidence of oral cancer has increased somewhat but significantly for both sexes from the middle of the 2000s to the most recent National Cancer Institute survey from 2015 to 2019 (SEER, 2023).

Over the past three decades, both sexes' oral cancer incidence rates for Blacks have considerably dropped (SEER, 2023). Incidence rates for oral cancer are greater

in White men than in Hispanic and Black men, and oral cancer incidence rates rise with age (Ellington, 2020). Increasing rates among whites and declining rates among blacks and Hispanics contributed to the overall trend of rising oral cancer rates. The incidence rates of malignancies of the base of the tongue, anterior tongue, gums, tonsils, oropharynx, and other oral cavity and pharynx all increased between 2007 and 2016, but the cancers of the lip, mouth floor, soft palate and uvula, hard palate, hypopharynx, and nasopharynx all saw a drop in incidence rates (Ellington, 2020).

Alcohol intake, cigarette use, and human papillomavirus (HPV) infection are the main etiological factors of oral cancer. A weaker immune system, inadequate nutrition, genetic disorders, prior mouth cancer diagnosis, prior oral cancer treatment, and family history of oral or other types of cancer are additional risk factors. It is also thought that genetic predisposition and inflammation are crucial contributors to the emergence of oral cancer (Wong, 2018).

The stage of cancer, the location of the tumor, and the severity of the disease all have an impact on the survival rates for oral cancer. From 2006 to 2007, the overall 5-year survival rate for Americans with oral or oropharyngeal cancer was 67%; the rates were 51% for Black Americans and 69% for White Americans. According to the American Cancer Society, surgery, radiation therapy, and chemoradiation are effective therapies for persons with oral cavity cancer in stages 1 and 2 (National Cancer, 2023). Based on how far cancer has spread, the Surveillance, Epidemiology, and End Results SEER database tracks the 5-year relative survival rates for oral cavity and oropharyngeal malignancies in the United States. As people get older, their chances of surviving malignancies of the mouth and pharynx decrease. Overall, survival rates only take into account cancer's stage at the time of diagnosis;

additional elements that may affect prognosis include the patient's age and the tumor's location (National Cancer, 2023).

Artificial intelligence (AI) is a term that refers to the use of machines and technology to carry out human-like tasks. AI models use the current set of observations to forecast future events (Ahmed, 2021). "Barr and Feigenbaum" defined AI as the branch of computer science that focuses on creating intelligent computer systems that display traits we associate with intelligence in human behavior, such as language comprehension, learning, reasoning, and problem-solving, to name a few (Barr A., 1981). Machine learning and its related topics, including deep learning, cognitive computing, natural language processing, robotics, expert systems, and fuzzy logic, are subcategories of AI. The main objective of machine learning is to enable automated learning without human judgment (Alhazmi, 2020). Machine learning has the capacity to progressively identify patterns, collect data, go through automated training based on data input, particularly complex non-homogenous data, and eventually produce clinical predictions with little to no human involvement (Chu, 2020).

To forecast bad clinical outcomes following Oral Squamous Cell Carcinoma (OSCC) treatment, some key data such as demographic, clinicopathological, therapeutic, and biomolecular data have all been used to construct clinical decision-making tools, such as statistical regression models and prognostic nomograms. Unfortunately, due to issues with data quality and low prediction accuracy, such approaches have only found limited acceptance in modern clinical practice. Recently, oncology research has used machine learning algorithms to construct analytical models to improve clinical outcome predictions (Adeoye, 2021).

For instance, Kwak et al. (2021) analyze the lymph node metastasis (LNM) in early Tumor (T) classification for oral squamous cell carcinoma using the SEER Database and machine learning predictive models. The result showed that comparing six machine learning-based prediction models, extreme Gradient Boosting (XGBoost) achieved an AUC of 0.956 for lymph node metastasis in early T classification OSCC patients. The permutation importance analysis revealed that tumor size was the most important factor in predicting metastasis. The study concluded that a reproducible and streamlined machine learning-based predictive model could assist clinicians in making informed decisions about treatment plans for early T classification OSCC patients.

## CHAPTER 2

### LITERATURE REVIEW

#### 2.1 Introduction

Oral cancer, particularly squamous cell carcinoma emerging from the oral mucosal lining contributes to half of the annual global mortality ascribed to head and neck cancer, with cancer-related deaths associated with aggressive source tumors and advanced-stage disease at clinical presentation. (Chu, 2020).

Machine learning, a subset of artificial intelligence, is the process by which computers learn from data and develop insights to anticipate new data based on the acquired information. Thus, current advancements in machine learning and neural networks could play a crucial role in assisting clinical decisions for patients using existing demographic, radiology, pathology, and fractional data. (Kwak, 2021). This section reviews three articles on the machine learning approach (Kwak et al, Peng et al and Chu et al) for predicting oral cancer-related mortality.

Peng et al. (2022) predicted the overall survival (OS) and cancer-specific survival (CSS) rates of HNC patients in each tumor site using five machine learning algorithms, and their prediction performance was compared with the Cox regression model. The study also evaluated each variable's significance in the optimal prediction model.

Peng et al. (2022) utilized Head and Neck Cancer HNC data from the SEER Program (Surveillance, Epidemiology, and End Results). The third edition of the International Classification of Oncology Diseases (ICD-O-3) was used to diagnose HNC in the patients in the sample set between 1974 and 2016. Data were extracted

using the SEER\*Stat software (version 3.8.8). The following were the exclusion criteria: T classification of 189 patients was TO or their Tumor, metastasis, and node (TM N) classifications were unknown for 186,398 patients. The primary sites for 324,648 patients were not the HNC mentioned above. About 8000+ patients were with incomplete follow-up information. Patients with missing data regarding age, sex, race, marital status, insurance, surgery, radiation, or chemotherapy were 7,762. The 7th edition of the American Joint Committee on Cancer Tumor, Node, and Metastasis (AJCC TNM) staging system served as the basis for the models that were built in the study, and examination of the significance of each variable in the superior prediction model (Peng et al., 2022). The SEER database's tumor sites for HNC allowed to categorize HNC patients into seven groups: major salivary gland cancer (MSGC), nasopharyngeal cancer (NPC), nasal cancer (NC), oral cavity cancer (OCC), hypopharyngeal cancer (HPC), laryngeal cancer (LC), nasal cancer (NC), and oropharyngeal cancer (OPC). Predictive variables included clinicodemographic information such as age, sex, race, insurance, state, year of diagnosis, second cancer, number of tumors, size, extension, lymph metastasis, histological type, grade, stage (general), T stage, N stage, M stage, surgery, radiation, chemotherapy, and HPV status. The OS and CSS variables were the survival outcomes, and the survival months variable was used to measure the length of OS and CSS. Patients who passed away during the follow-up period for any reason were classified as OS, while those who passed away from HNC were classified as CSS. There were five machine learning models applied. Specifically, support vector machines (SVMs), neural networks (NN), random forests (RF), gradient boosting decision trees (GBDT), and deep learning (DL). A Python package was utilized for experimentation and

implementation with all survival models. To implement the COX, RF, and SVM, respectively, Cox net Survival Analysis, Random Survival Forest, and Fast Survival SVM were used. To assess the prediction performance of the models, a IO-fold cross-validation schema was used. The concordance index (C-index) and integrated Brier score (IBS), were essential for censored data, as the primary evaluation index. The average of each metric was then reported overall 10 folds. The data was analyzed using the R language (4.0.4) and Python (3.6).

Using the 7th edition of the AJCC TNM staging system, Peng et al. (2022) found that 64,226 eligible cases in total were identified. Of these, 57,803 cases were randomly assigned to the training cohort and 6,423 cases were randomly assigned to the testing cohort (Peng et al., 2022). Training and validation for the training cohort was used for 52,023 cases and 5,780 cases respectively.. The initial features of patients with primary hematologic cancer are categorized by tumor site. The oropharynx was the most frequently involved site (36%, OPC), and two-thirds of OPC patients (73%) tested positive for HPV. The larynx (24%, LC) and the oral cavity (20%, OCC) were the next most common sites. Furthermore, 80% of NPC patients had chemotherapy, compared to 84% of MSGC and 81% of OCC patients who underwent surgery as their primary course of treatment. The duration of survival was 30.9 months on average (range: 1-83 months). When taking into account a wide range of indicators, the RF algorithm generally performed better than the other four machine learning algorithms and the conventional Cox regression model. Primary outcomes were calculated, such as IBS and C-index. Secondary outcomes were noted, such as the time-dependent ROC's AUC value. The RF model's time-dependent ROC curves and AUC values were higher than those of other models for every HNC tumor site.

Based on the HPV status of 12,271 OPC patients, a subgroup analysis was carried out. In both the testing and validation cohorts, the RF model consistently outperformed other algorithms in prognosis prediction. The models were retrained and retested before external validation because this external database had several variables missing that were present in the original database. Throughout this external database, the RF model consistently outperformed other models in terms of prognostic predictive effects (Peng et al., 2022).

Chu et al. (2020) assessed the potential of supervised machine-learning models to predict disease outcomes by taking into account a well-characterized patient cohort. Chu et al. (2020) carried out a retrospective assessment using data from the Hospital Authority Clinical Management System (HA CMS) of OSCC patients treated at Queen Mary Hospital in Hong Kong during a 19- year period, from October 1, 2000, to October 1, 2019.

Thirty-four demographic, clinicopathological, and lifestyle parameters were chosen as prognostic features to be included in the prediction models (Chu et al., 2020). These factors were taken from the database based on their correlation with the risk of progressive disease. Age, sex, date of diagnosis, status (living or dead) at the time of data retrieval, history of cancer in the past, alcohol and/or tobacco use, human papillomavirus (HPV), and Epstein-Barr virus (EBV) status were among the patient's demographic details. Tumor site, grading, histological features of tumor invasiveness, resection margin status, pTNM classification, disease staging, and, when applicable, the use of adjuvant chemo-radiotherapy regimens and/or cervical lymph node dissection were all documented in the clinicopathological data. Either a progressing disease, as shown by distant metastases developing or a loco-regional tumor

recurrence, or no disease at all was noted as the outcome. From the date of the initial diagnosis to the patient's death or the most recent clinic follow-up, the overall survival was calculated. Models of prediction. The four commonly used models for outcome prediction-linear regression (LR), decision tree (DT), support vector machine (SVM), and k-nearest neighbors (KNN)-were constructed using MATLAB R2020a (MathWorks, Inc.), a mathematical programming platform that facilitates data graphic illustration and evaluation. The models were developed using the 34 prognostic features (predictors) and the existence of progressive disease (outcome). To prevent overfitting, 15-fold cross-validation was employed to assess the models. Prior to constructing the models, two techniques were employed to explore the possibility of data reduction improving performance: principal component analysis (PCA) for dimensionality reduction and highlighting correlated variables, and bivariate analysis for identifying prognostic features that were positively correlated with the outcome ( $P < .05$ ). The receiver operating characteristic (ROC) and area under the curve (AUC) calculations were used to evaluate the diagnostic capacity of the models, which were validated for accuracy, sensitivity (true positive), and specificity (true negative). Predictive performance employing all 34 prognostic features for each model was compared to results from bivariate analysis and PCA (Chu et al., 2020).

Chu et al. (2020) discovered that 467 OSCC patients in total were found in the HA CMS database, and their complete clinicopathological and demographic information was compiled. In the end, 59 patients were not included in the study analysis because clinicopathological data were not available (43), or because patients who were listed as alive did not show up for their most recent clinic assessments (16). Thus, the machine-learning models were loaded with data from 408 OSCC patients

(244 males and 164 females); at the time of data retrieval, 151 (37%) had passed away and 131 (32%) showed signs of progressive disease. 13 prognostic factors were found to be positively predictive of the development of progressive disease in the bivariate analysis, whereas the PCA used 16-34 components, depending on the model. The highest accuracy of 70.83% (AUC 0.68) was obtained by PCA for the LR model reduced to 18 components, as compared to using all 34 predictive features or 13 selected by bivariate analysis. Additionally, PCA had the best specificity (88.81%) while the 34-feature model had a higher sensitivity (35.11%). Better than both bivariate analysis and 19-component PCA, the 34-feature DT model achieved 70.59% accuracy (AUC 0.67), with a sensitivity of 41.98% and specificity of 84.12%. The accuracy of the 34-feature SVM model and the 34-component PCA were identical (69.85%, AUC 0.68). After applying PCA, the sensitivity rose from 24.43% to 25.95%, but the specificity decreased by 0.73%. Although specificity increased to 93.86%, bivariate analysis decreased accuracy and sensitivity to 68.63% and 15.27%, respectively. All three of the SMV models had good specificity overall—all three scoring above 90%. With a 69.36% accuracy rate (AUC 0.71), 35.11% sensitivity, and 85.56% specificity, the KNN model performed well. While sensitivity was the same at 25.95%, PCA (16 components) outperformed the 34-feature model in terms of accuracy (68.38% vs. 66.42%) and specificity (88.45% vs. 85.56%). The most successful model in detecting "true positive" progressive disease was the DT model with 34 prognostic features; it achieved 70.59% accuracy, 41.98% sensitivity, and 84.12% specificity. Across all models, specificity ranged from 79.42% to 93.86%, while sensitivity ranged from 15.27% to 41.98% (Chu et al., 2020).

Kwak et al. (2021) developed a new model utilizing six machine learning classifiers that used fundamental clinical and histological criteria from public SEER data to predict LNM in early T classification OSCC patients.

Kwak et al. (2021) used information from the National Cancer Institute (NCI) SEER Program database, which was established to gather information on cancer incidence, survival, and prevalence from 17 population-based cancer registries in the United States. This database currently includes information on 34.6% of the country's population. All patients diagnosed with T1 and T2 OSCC between 2004 and 2016 had data from their records examined using the SEER 1975-2016 database (published on April 15, 2019). Using SEER disease codes, clinical demographic data, such as age at diagnosis, sex, race, and tumor characteristics such as size, grade, and American Joint Committee on Cancer 7th TNM stages were extracted. Using ICD-O-3 histology codes, the histologic types were restricted to squamous cell carcinoma. The World Health Organization's four-tier classification system-well-differentiated, moderately differentiated, poorly differentiated, and undifferentiated-was applied to the grading of tumor differentiation. In their study, patients with early T-stage OSCC were included; patients with any missing data were not included. LNM in early T classification OSCC patients was predicted using six machine learning based models, which were chosen as the prevalent and up-to-date predictive model type. The logistic regression model (LR), support vector machine (SVM), classification and regression trees model (CART), XGBoost (XGB) model, random forest (RF) model, and k-nearest neighbor algorithm (kNN) are the six models that were used in the study. In order to improve the classifier performance for small classes in an unbalanced dataset, a random oversampling technique was employed. While the ML algorithm was being

trained, k-fold cross-validation was carried out, and grid search was used to exhaust all possible parameter combinations. The number of cases correctly classified as TP (true positive) and the number of cases incorrectly classified as FP (false positive) were used to compare the performance of various algorithms. True negatives (TN) represent the number of cases that were correctly classified as non-metastasis, while false negatives (FN) represent the number of cases that were incorrectly classified in this manner. Sensitivity (recall), specificity, accuracy, average precision (AP), F1 score, and Matthew's correlation coefficient were determined using TP, FP, TN, and FN. area under the curve (AUC) of the confusion matrix was used as a performance indicator for each algorithm. A permutation feature importance score using the test and train set was obtained, which allowed for assessing the most crucial variables in the model prediction. Using the NCI's SEER\* Stat software (Surveillance Research Program, National Cancer Institute SEER\* Stat software, version 8.3.6), all data were taken out of the SEER database. R statistical software (version 3.6.0) and Python (version 3.6.9) were used for all analyses. Using the Student's t-test and Pearson chi-square test, the baseline characteristics of the two groups were compared based on whether metastasis was present (Kwak et al., 2021).

Using the method described above, in the study conducted by Kwak et al. (2021), a total of 24,547 patients with OSCCs were gathered between 2004 and 2016, of which approximately 7019 patients were omitted from the investigation due to their advanced T classification (T3 or T4) diagnosis following the exclusion of 650 patients because there was insufficient data on the nodal, tumor grade, and tumor size. Using CART classification, regression trees model; kNN, k-nearest neighbor algorithm; LR, logistic regression model; OSCC, oral squamous cell carcinoma; RF, random forest;

SVM, support vector machine; and XGB, XGBoost 16878 patients with a diagnosis of early T classification OSCC (T1 and T2) were examined. Six independent prognostic variables were included in the model: T classification, tumor grade, tumor size, age at diagnosis, sex, race, and primary site. The patients that were included were split into groups based on whether they had metastases (2731 patients, 16.2%) or not (14 147 patients, 83.8%). Patients with LNM had a mean age that was significantly lower than patients without metastases ( $p < 0.001$ ). In comparison to the non-metastasis group, the proportion of males in the metastatic group was significantly higher ( $p = 0.006$ ). The group with metastasis had a significantly higher proportion of metastasis in the tongue, floor of mouth (FOM), and buccal regions ( $p < 0.001$ ). Between the two groups, there was a notable difference in the distribution of races. Out of the six models, the XGB model had the highest AUC value (0.956), while LR had the lowest AUC value (0.768). With an AP value of 0.946, XGB had the highest of the six models' AP values, all of which were 0.78 or higher. The XGB, CART, and RF models had comparatively good sensitivity and specificity out of the six models. The LR and SVM models, on the other hand, displayed poor sensitivity and specificity. The majority of models had accuracy values above 70; the XGB model had the highest accuracy, at 88.879, while the SVM model had the lowest, at 69.666. The XGB model also fared best in terms of the F1 score and Matthew's correlation coefficient value. Permutational importance factors for early T classification oral cancer metastasis in the test set were used to quantify the variable importance for XGB. The features included in the machine learning models that were found to be most significant were tumor size and T classification (Kwak et al., 2021).

## 2.2 Integration of the individual articles

Each of the three studies evaluated the effectiveness of machine learning models for forecasting oral cancer mortality. However, the participant sampling and data sources were quite different. Peng et al (2022), Kwak et al (2021), and Chu et al (2020) used secondary data. Chu et al (2020) used records from the Hospital Authority Clinical Management System, whereas Peng et al (2022) and Kwak et al (2021) used data from the National Cancer Institute (NCI) SEER Program database, which was created to collect cancer incidence, prevalence, and survival data from U.S. cancer registries. Chu et al. (2020) employed patient demographic data, such as age, sex, date of diagnosis, status (alive or dead) at the time of data retrieval, and information about smoking, alcohol, human papillomavirus (HPV), and Epstein-Barr virus (EBV) status, and clinicopathological data recorded tumor site, grading, histopathological characteristics of tumor invasiveness, resection margin status, pTNM classification, disease staging, and, when appropriate, the use of cervical lymph node dissection and/or adjuvant chemo-radiotherapy regimens. Meanwhile, Kwak et al. (2021) extracted clinical demographic data, including age at diagnosis, sex, race, histologic types, and tumor information about the primary site, size, grade, and American Joint Committee on Cancer 7th TNM stages. Peng et al (2022) assessed clinicodemographic data such as age, sex, race, marital status, insurance, state (the area of the cases), year of diagnosis, second cancer, tumor number, tumor size, extension, lymph metastasis, histological type, grade, stage (overall), T stage, N stage, M stage, surgery, radiation, chemotherapy and HPV status were included as prediction variables.

Kwak et al (2021) used six machine learning models including logistic regression model (LR), support vector machine (SVM), classification and regression trees model (CART), XGBoost (XGB) model, random forest (RF) model, and k-nearest neighbor algorithm (kNN) whereas Peng et al (2022) used Random forest (RF), Gradient boosting decision tree (GBDT), Support vector machines (SVMs), Neural network (NN), and Deep learning (DL), however, Chu et al (2020) employed linear regression (LR), decision tree (DT), support vector machine (SVM) and k-nearest neighbours (KNN) models.

Kwak et al (2021) and Peng et al (2022) performed analysis using Python (version 3.6.9) and R statistical software (version 3.6.0) whereas Chu et al executed evaluation using MATLAB R2020a (MathWorks, Inc.), a mathematical programming platform facilitating data plotting and analysis.

Chu et al. (2020) prevented overfitting assessing models using 15-fold cross-validation. In contrast, Peng et al (2022) and Kwak et al (2021) utilized a 10-fold cross-validation schema to determine the average of each metric over all 10 folds, and Peng et al (2022) also performed the nonparametric MissForest approach to improve the data's fit to the models.

Kwak et al (2021) used the student's t-test and Pearson chi-square test, the baseline characteristics of the two groups were compared based on whether metastasis was present.  $P < 0.05$  on both sides was regarded as statistically significant, whereas Chu et al (2020) performed bivariate analysis using IBM SPSS for Windows 10 version 25 was employed to detect prognostic features that had a positive correlation with the outcome ( $P < .05$ ). However, Peng et al (2022) did not employ any kind of bivariate analysis.

Chu et al (2020) utilized receiver operating characteristic (ROC) and area under the curve (AUC) calculations accuracy, sensitivity, and specificity to predictive performance employing all 34 prognostic characteristics for each model, while Kwak et al. (2021) assessed model performance using sensitivity (recall), specificity, accuracy, average precision (AP), F1 score, and Matthews correlation coefficient and area under the curve (AUC). However, Peng et al (2022) employed the concordance index (C-index), and the integrated Brier score (IBS), as the primary evaluation indexes to assess the prediction performance of the models and time-dependent receiver operating characteristic (ROC) curves were implemented to further estimate the models, with accuracy and the area under the curve (AUC) serving as the secondary evaluation metrics.

In terms of sample size and data source, Chu et al. (2020) had the lowest sample size of 467 OSCC patients, and data was extracted from the HA CMS database. This was followed by Kwak et al. (2021) with a sample size of 24 547 patients diagnosed with OSCCs between 2004 and 2016 from the SEER program database. Peng et al (2022) had the highest sample size of the three studies, with 64,226 eligible cases included in their study following the 7th edition of the AJCC TNM staging method.

Chu et al. (2020), after comparing four ML models (DT vs LR vs SVM vs KNN), the DT model with 34 prognostic variables seemed to be the most effective at detecting "true positive" progressive disease, with accuracy (70.59% vs 67.89% vs 69.85% vs 66.42%, respectively), AUC (0.67 vs 0.68 vs 0.68 vs 0.69, respectively), sensitivity (41.98% vs 35.11% vs 24.43% vs 25.95%, respectively), and specificity (84.12% vs 83.39% vs 91.34% vs 85.56%, respectively), whereas according to Peng

et al (2022), after comparing five machine learning models (RF vs GBDT vs DL vs NN vs SVM), it was concluded that the RF algorithm was superior to other four machine learning models and the traditional Cox regression model when various indicators were comprehensively considered. The primary outcomes for oral cavity cancer in terms of overall survival rate were determined by C-index (80.130+/- 0.4800 vs 78.070+/- 0.5500 vs 75.740+/- 0.7500 vs 51.570+/- 1.1200 vs 46.790+/- 0.7200, respectively), and IBS (14.510+/- 0.2400 vs 17.910+/- 0.3900 vs 17.270+/- 0.6300 vs no values for NN & SVM; respectively). The secondary outcome was determined by the AUC value which was 0.1684 for SVM; 0.8087 for RF; 0.7895 for Cox and 0.7725 for DL. In contrast, Kwak et al (2021) post-evaluation of six machine learning models i.e. LR vs XGB vs kNN vs CART vs SVM vs RF wherein XGB was superior to other models with AUC ( 0.768 vs 0.956 vs 0.842 vs 0.843 vs 0.769 vs 0.896, respectively), sensitivity ( 68.339% vs 91.590% vs 68.362% vs 84.899% vs 68.103% vs 83.415% ), specificity ( 72.973% vs 86.168% vs 86.592% vs 83.435% vs 71.229% vs 83.46%, respectively), AP ( 0.785 vs 0.946 vs 0.870 vs 0.880 vs 0.786 vs 0.895, respectively), F1 score (0.71 vs 0.89 vs 0.77 vs 0.84 vs 0.70 vs 0.83), and Matthews correlation coefficient ( 0.413 vs 0.778 vs 0.558 vs 0.683 vs 0.393 vs 0.666, respectively).

### **2.3 Discussion of reviewed articles:**

Three studies have assessed the ability of supervised machine learning models to predict oral cancer-related mortality. These studies were generally similar based on the fact that they were retrospective studies that utilized secondary data for the study.

However, there were substantial methodological differences across the studies. For example, two studies derived the dataset from the SEER database (Kwak,2021; Peng,2022) while the other study (Chu,2020) performed analysis using the Hospital Authority Clinical Management System. Additionally, the studies varied based on sample size, the variables assessed, and statistical analysis.

All the participants in the studies had experienced oral cancer. Oral cancer prevalence is based on numerous predictor variables. According to Peng et al. (2022), the oral cavity and larynx were the next most often implicated sites, with two-thirds of OPC patients being HPV positive. The oropharynx was the most usually involved site. The RF model yielded consistent findings from both internal and external validation, and it was the most accurate in predicting the prognosis of HNC. The RF model yielded the greatest prognostic prediction in each HNC subtype, with age and tumor size exhibiting the largest prognostic effects. The results were further validated using both internal (the 6th edition of the AJCC TNM staging system cohort) and external (the TCGA cohort) validations, and the generalizability of the models was constantly maintained. SVM, NN, and DL as single learning models are outperformed by RF and GBDT, especially RF. Compared to other machine learning models, RF performs better in generalization and prediction when presented with noisy data and missing values. The highest C-index was produced by the RF model. The RF model's time-dependent ROC curves and AUC values were greater than those of other models for each of the HNC tumor sites. The RF model consistently outperformed the other models in the performance prediction of a subgroup analysis of 12,271 OPC patients depending on HPV status, while Kwak et al. (2021) developed six models that incorporated the clinical and pathologic parameters of 16,878 OSCC patients with T1,

T2, and predicted LNM. Out of the six models, the XGB model maintained its high AUCs, and the models with the highest accuracy predictions were the XGB, CART, and RF models. Chu et al. (2020) proposed that the 34 prognostic indicators in the DT model seemed to have a higher specificity for diagnosing "true positive" progressing disease. Across all models, specificity was generally far higher than sensitivity. All 12 models in this analysis performed fairly, but DT, which employed 34 prognostic characteristics, was the best at predicting the course of OSCC. Despite their numerous shortcomings, the research had a number of advantages. Large participant numbers were used in the investigations, which is ideal to guarantee the validity and generalizability of the findings. The findings were mostly in line with the majority of other research that have been published in American and international literature. Overall, the analyses made sense given the data they had, and they did not draw the wrong conclusions about causality.

## **2.4 Conclusion**

The previous studies conclude that the machine learning models have produced encouraging findings in terms of forecasting the course of Oral Squamous Cell Carcinoma. The research findings indicate that the application of machine learning and artificial intelligence techniques with huge databases can significantly contribute to the development of treatment plans by enhancing the quantification of individual patient metastatic risk estimates before therapy. This is intended to detect progressive illness at the earliest and improve knowledge of the biological processes underlying aggressive tumor activity. So, to sum up, all three studies claimed that

artificial intelligence could help healthcare professionals make more informed decisions and personalized treatment plans, predict patient outcomes, and rationalize interventions by using digital health information.

## **2.5 Gaps in the studies reviewed:**

The aforementioned studies have several drawbacks. Kwak et al. (2021) did not use a comprehensive and detailed version of prognostic features as Tumor Stage grading, treatment options, diagnostic histology, N classification, and M classification were not included. Peng et al. (2022) did not consider metrics such as accuracy, specificity, sensitivity, brier score, or kappa statistics. Chu et al. (2020) did not include the prognostic features such as treatment options, and diagnostic histology which was later explored in our study.

These limitations were addressed in our study. Compared with other studies, our study included the most detailed version of the prognostic features including sociodemographics, histopathological characteristics and treatment options, Tumor Stage grading, treatment options, diagnostic histology, N classification, and M classification with the most recently available SEER dataset from 1992 to 2021 released on April 2023. Additionally, our study included statistical metrics like accuracy, specificity, sensitivity, brier score, and kappa statistics.

Based on the gaps in the previous studies, we proposed to test whether the benefits of machine learning are replicated in the predictive model for oral cancer mortality. Specifically, the objectives for our study will be to :

To predict oral cancer-related mortality among adults in the United States using the Machine Learning approach.

To identify the predictors of oral cancer-related mortality using the Machine Learning approach.

## CHAPTER 3

### MATERIAL AND METHODS

#### 3.1 Data Source

Data was extracted for a sample size of about 8000 oral cancer patients from the Surveillance, Epidemiology, and End Results Program ([www.seer.cancer.gov](http://www.seer.cancer.gov)] SEER\*Stat Database: Incidence - SEER Research Plus Data, 12 Registries, Nov 2022 Sub (1992-2020) - Linked to County Attributes - Time-Dependent (1990-2021) Income/Rurality, 1969-2021 Counties, National Cancer Institute, DCPS, Surveillance Research Program, released April 2023, based on the November 2022 submission. Institutional review board (IRB) approval was waived because SEER is a deidentified governmental database. Data were extracted and reported following the SEER database user agreement.

The outcome was noted as "Death Class" (SEER cause-specific death classification) indicating Oral Cancer- Related Mortality coded as either Died from Oral Cancer or Alive/Died from other causes. The variable in the SEER data for estimating cause-specific survival probability due to 'cancer' or due to 'other causes.' The idea was to use the variables independently to estimate survival of specified causes of death (e.g. cancer, non-cancer). The 'SEER cause-specific death classification' variable is used to obtain cancer-specific survival probability for a given cohort of cancer patients. In the variable, (SEER cause-specific death classification) deaths attributed to the cancer of interest are treated as events, and deaths from other causes are treated as censored observations.

### **3.2 Prognostic Features**

A series of 38 sociodemographic, clinicopathological & histological, and Treatment alternatives, were extracted from the SEER database in view of their association with progressive disease risk and were selected as prognostic features to populate the prediction models.

The 37 prognostic features included Age Group, Gender, Year of Diagnosis, Race and Origin, Cancer Site, Diagnostic Confirmation, Clinical Grade, Laterality, Histology Behavior, Rare Tumors, Derived T Stage, Derived N Stage, Derived M Stage, Derived Stage Group, Scope of LN Surgery, Other Surgical Procedures, Surgery and Radiation Sequence, Radiation Type, Systemic Therapy and Surgery Sequence, Lymph Node Size, Bone Metastasis at Diagnosis, Brain Metastasis at Diagnosis, Liver Metastasis at Diagnosis, Lung Metastasis at Diagnosis, Distant Lymph Node Metastasis at Diagnosis, Other Distant Metastasis at Diagnosis, Cause-Specific Death Classification, Survival Months, Sequence Number, First Malignant Primary, Primary by International Rules, Year of Follow-up Recode, Year of Death Recode, Detailed Cancer Site, Reporting Source, Marital Status, Median Household Income, and Rural-Urban Continuum Code. The power determination/sample size calculation was not possible as the study included numerous variables.

### **3.3 Detailed Definition of Independent Variables**

**AGE:** The age recode variable is based on Age at Diagnosis (single-year ages). The groupings used in the age recode variable are determined by the age

groupings in the population data. This recode has 19 age groups in the age recode variable(< 1 year, 1-4 years, 5-9 years, ..., 85+ years).

**SEX:** The gender of the patient variable is used to link to the correct populations for males and females when calculating sex-specific rates.

**YOD (ear of diagnosis):** The year of diagnosis is the year the tumor was first diagnosed by a recognized medical practitioner, whether clinically or microscopically confirmed.

**RACE (Race and origin):** SEER provides the Race and Origin Recode variable with the following values: Non-Hispanic White, Non-Hispanic Black, Non-Hispanic American Indian/Alaska Native, Non-Hispanic Asian or Pacific Islander, Hispanic (All races), and Non- Hispanic Unknown Race.

**ICD-O-3 (Site recode ICD-O-3 WHO 2008):** SEER site recode variables are based on the primary site and histology data fields submitted to SEER by the registries. The site recode variables define the major cancer site/histology groups that are commonly used in SEER's reporting of cancer incidence data and published statistics. Based on ICD-0-3, updated for Hematopoietic codes based on WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues (2008). (Campo, 2008)The subcategories for the variable are Floor of the mouth, gums and other sites in the mouth, lip, salivary gland, and tongue.

**DIAGNOSTIC (Diagnostic Confirmation):** This data item records the best method used to confirm the presence of the cancer being reported. The best method could occur at any time throughout the entire course of the disease. It is not limited to the confirmation at the time of initial diagnosis. The subtypes include clinical diagnosis only, positive exfoliative cytology, no positive histology, positive histology,

positive microscopic confirm/method not specified, and radiography without microscopic confirm.

**GRADE (Grade Clinical 2018+):** This data item records the grade of a solid primary tumor before any treatment (surgical resection or initiation of any treatment including neoadjuvant).

- Rationale: Grade is a measure of the aggressiveness of the tumor. Grade and cell type are important prognostic indicators for many cancers. For some sites, the grade is required to assign the clinical stage group for those cases that are eligible AJCC staging, the recommended grading system is specified in the AJCC Chapter. The AJCC Chapter- specific grading systems (codes 1-5) take priority over the generic grade definitions (codes A-E, L, H, 9). For those cases that are not eligible for AJCC staging, if the recommended grading system is not documented, the generic grade definitions would apply.

For solid tumors diagnosed in 2018 and forward, grade will be collected in three different data items, Grade Clinical, Grade Pathological, and Grade Post Therapy, and the codes and coding instructions will depend on the type of cancer.

- Code Grade Description: 1 stands for Site-specific grade system category, 2 stands for Site- specific grade system category, 3 stands for Site-specific grade system category, 4 stands for Site-specific grade system category, 5 stands for Site-specific grade system category, 8 stands for Not applicable (Hematopoietic neoplasms only) 9 stands for Grade cannot be assessed; Unknown.

A stands for Well differentiated, B stands for Moderately differentiated, C stands for Poorly differentiated, D stands for Undifferentiated and anaplastic, and E stands for Site-specific grade system category.

**LAT (Laterality):** Laterality describes the side of a paired organ or side of the body on which the reportable tumor originated. Determine whether laterality should be coded for each primary.

The sub-categories for laterality include bilateral primary tumor, primary tumor on the left side, unpaired tumor site, unilateral with unspecified side, paired tumor site, and primary tumor on the right side.

**ICD-0-3-HIS (ICD-0-3 Histology & malignant behavior):** Labeled version of ICD-O-3 values for malignant tumors. All non-malignant tumors are grouped into one value.

- The sub-categories are: 8000/3: Neoplasm, malignant; 8010/3: Carcinoma, NOS; 8013/3: Large cell neuroendocrine carcinoma; 8041/3: Small cell carcinoma, NOS; 8046/3: Non-small cell carcinoma; 8051/3: Verrucous carcinoma, NOS; 8052/3: Papillary squamous cell carcinoma; 8070/3: Squamous cell carcinoma, NOS; 8071/3: Squamous cell carcinoma, keratinizing, NOS; 8072/3: Squamous cell carcinoma, large cell, nonkeratinizing, NOS; 8074/3: Squamous cell carcinoma, spindle cell; 8075/3: Squamous cell carcinoma, adenoid; 8082/3: Lymphoepithelial carcinoma; 8083/3: Basaloid squamous cell carcinoma; 8085/3: Squamous cell carcinoma, HPV-positive; 8086/3: Squamous cell carcinoma, HPV-negative; 8090/3: Basal cell carcinoma, NOS; 8140/3: Adenocarcinoma, NOS; 8147/3: Basal cell adenocarcinoma; 8200/3: Adenoid cystic carcinoma;

8246/3: Neuroendocrine carcinoma, NOS; 8290/3: Oxyphilic adenocarcinoma; 8310/3: Clear cell adenocarcinoma, NOS; 8407/3: Sclerosing sweat duct carcinoma; 8410/3: Sebaceous adenocarcinoma; 8430/3: Mucoepidermoid carcinoma; 8480/3: Mucinous adenocarcinoma; 8500/3: Infiltrating duct carcinoma, NOS; 8502/3: Secretory carcinoma of breast; 8525/3: Polymorphous low grade adenocarcinoma; 8550/3: Acinar cell carcinoma; 8560/3: Adenosquamous carcinoma; 8562/3: Epithelial-myoeithelial carcinoma; 8720/3: Malignant melanoma, NOS; 8746/3: Mucosal lentiginous melanoma; 8772/3: Spindle cell melanoma, NOS; 8941/3: Carcinoma in pleomorphic adenoma; 8980/3: Carcinosarcoma, NOS; 8982/3: Malignant myoepithelioma.

**TUM (Site and Rare cancer type):** Cancer type grouping to define clinically relevant, histologically defined rare cancers. The list of rare cancers was developed by the Surveillance of Rare Cancer in Europe (RARECARE) to include clinically relevant, histologically defined rare cancers with an annual crude incidence rate smaller than 6 per 100,000. The original list of rare cancers was recently revised and applied to the European data and the SEER data and is now available in SEER\*Stat. The subtypes include epithelial tumors of major salivary glands, Salivary gland-type tumors of the head and neck, malignant melanoma of mucosa and extracutaneous layer, squamous cell carcinoma with variants of the oropharynx, other epithelial tumors of oropharynx, neuroendocrine carcinomas of other sites, squamous cell carcinoma of the oral cavity, squamous cell carcinoma with variants of the lip.

**T\_stage (EOD T classification 2018+):** refers to a variable or data field within the SEER database. Specifically, it appears to be related to the Extent of

Disease (EOD) coding system for cancer, specifically for the T (Tumor) category, and it is specified for the year 2018.

- T0: Indicates that there is no evidence of a primary tumor. In some cases, this means that the tumor cannot be found despite thorough examination.
- T1: Represents a small tumor that is typically confined to the organ or tissue of origin and is often associated with a better prognosis.
- T2: Signifies a larger tumor than T1 but is still generally confined to the organ or tissue of origin.
- T3: Indicates a larger tumor that may have started to invade nearby tissues or structures.
- T4: Represents a tumor that has further invaded adjacent tissues, organs, or structures. T4 is often subdivided into T4a and T4b, with T4a typically indicating limited invasion, and T4b indicating more extensive invasion.
- Tis: Denotes carcinoma in situ, which means the cancer cells are present but have not invaded beyond the layer of cells where they originated. It's considered an early stage.
- TX: Represents that the primary tumor cannot be assessed or is unknown.

**N\_stage(EOD N classification 2018+)**: refers to a variable or data field within the SEER database, specifically related to the Extent of Disease (EOD) coding system for cancer, specifically for the N (Regional Lymph Nodes) category, and it is specified for the year 2018.

- N0: Indicates that there is no evidence of regional lymph node involvement. This suggests that nearby lymph nodes do not contain cancer cells.
- N1: Represents the presence of cancer cells in nearby regional lymph nodes,

typically in the lymph nodes closest to the primary tumor.

- N2: Signifies regional lymph node involvement, and it can be further subdivided into N2a, N2b, and N2c to specify the extent and number of lymph nodes involved.
- N2a: Indicates involvement of lymph nodes that are a bit farther from the primary tumor.
- N2b: Signifies involvement of lymph nodes that are even farther from the primary tumor.
- N2c: Represents involvement of lymph nodes in a different drainage area from the primary tumor.
- N3: Indicates more extensive regional lymph node involvement than N2, and it can be further subdivided into N3a and N3b to specify the extent and location of lymph node involvement
- N3a: Denotes lymph node involvement in a distant drainage area from the primary tumor.
- N3b: Signifies involvement of lymph nodes near the primary tumor but with significant extension.
- NX: Represents that the regional lymph nodes cannot be assessed or are unknown.

**M\_stage(EOD M classification (2018+)):** Refers to a variable or data field within the SEER database, specifically related to the Extent of Disease (EOD) coding system for cancer, specifically for the M (Metastasis) category, and it is specified for the year 2018.

- MO: Indicates that there is no evidence of distant metastasis. In other

words, cancer has not spread to distant organs or tissues.

- **MI:** Represents the presence of distant metastasis. This means that cancer has spread to one or more distant sites in the body, such as distant organs or distant lymph nodes.

**Stage\_Group= EOD Stage Group classification (2018+):** Refers to a variable or data field within the SEER database, specifically related to the staging of cancer cases using the Extent of Disease (EOD) coding system for the year 2018. Rationale Derived EOD 2018 Stage Group can be used to evaluate disease spread at diagnosis, treatment patterns, and outcomes over time. Derived EOD 2018 Stage Group is only available at the central registry level.

**LN\_Surgery (Lymph node surgery):** This variable contains information related to the scope of regional lymph node surgery in cancer cases, starting from 2003 onwards. Scope of Regional Lymph Node Surgery describes the procedure of removal, biopsy, or aspiration of regional lymph nodes performed during the initial work-up or first course of therapy at all facilities. The subcategories are: 1 to 3 regional lymph nodes removed, 4 or more regional lymph nodes removed, Biopsy or aspiration of regional lymph node, NOS, None, Number of regional lymph nodes removed unknown, Sentinel lymph node biopsy, Sentinel node biopsy and lymph node removed different times, Sentinel node biopsy and lymph node removed same/unstated time, and Unknown or not applicable.

**SURG\_PROC (Distant/ Regional Surgical Removal):** This variable may contain data about other surgical procedures related to regional or distant metastases in cancer cases, starting from 2003 onwards. Surgical procedure of Other Site

describes the surgical removal of distant lymph node(s) or other tissue(s) or organ(s) beyond the primary site.

- The subcategories include Any combination of surgical procedures to other regional, distant lymph nodes, and/or distant site, Non-primary surgical procedure performed, Non- primary surgical procedure to distant lymph node(s), Non-primary surgical procedure to distant site, Non-primary surgical procedure to other regional sites, None; diagnosed at autopsy, and Unknown; death certificate only.

**SRS (Surgery Radiation Sequence):** This field records the order in which surgery and radiation therapies were administered for those patients who had both surgery and radiation.

- The sub-categories of the variable depicting surgery radiation sequence are as follows: No radiation and/or surgery as defined above, radiation before surgery, radiation after surgery, radiation both before and after surgery, intraoperative radiation, intraoperative radiation with other radiation given before and/or after surgery, Surgery both before and after radiation, Sequence unknown, but both surgery and radiation were given.

**RAD (Radiation):** This data item indicates the method of radiation therapy performed as part of the first course of treatment. The variable was created from RX Summ-Radiation, Item #s=1360, for 1975-2017 and Phase X Radiation Treatment Modality, Item #s=1506, 1516 and 1526, for 2018+. Unknown has been combined with None.

None/Unknown; diagnosed at autopsy, Beam radiation, Radioactive implants, Radioisotopes, Combination of 1 with 2 or 3, Radiation, NOS-method or source not

specified, Other radiation (1973-1987 cases only), Patient or patient's guardian refused radiation therapy, Radiation recommended, unknown if administered.

**SYS\_SURG\_SEQ (Systemic Therapy and Surgical Procedure Sequence):**

This data item records the sequencing of systemic therapy and surgical procedures given as part of the first course of treatment, including:

- No systemic therapy and/or surgical procedures; unknown if surgery and/or systemic therapy given, Systemic therapy before surgery, Systemic therapy after surgery, Systemic therapy both before and after surgery, Intraoperative systemic therapy, Intraoperative systemic therapy with other therapy administered before and/or after surgery, Surgery both before and after systemic therapy, Sequence unknown, but both surgery and systemic therapy given.

**LN\_SIZE (Lymph node size 2010):** This variable may contain data about the size of lymph nodes in cancer cases, starting from 2010 onwards.

**Bone\_Mets (Bone metastases):** These variables likely indicate the presence or absence of metastases (cancer spread) to bone at the time of diagnosis, starting from 2010 onwards.

**Brain\_Mets (Brain metastases):** These variables likely indicate the presence or absence of metastases (cancer spread) to the brain at the time of diagnosis, starting from 2010 onwards.

**Liver\_Mets (Liver metastases):** These variables likely indicate the presence or absence of metastases (cancer spread) to the liver at the time of diagnosis, starting from 2010 onwards.

**Lung\_Mets (Lung Metastases):** These variables likely indicate the presence or absence of metastases (cancer spread) to the lung at the time of diagnosis, starting from 2010 onwards.

**Dist\_LN\_Mets (Lymphnode Metastases):** This variable indicates the presence or absence of distant lymph node metastases at the time of diagnosis, starting from 2016 onwards.

**Other\_Mets (Metastases to Other Locations):** This variable contains information about metastases to other locations not covered by the previous variables, starting from 2016 onwards.

**Survival\_Flag (Survival months):** This variable indicates whether certain conditions were met for calculating survival months. The flag will enable analysts to easily select a subset of cases:

- Complete dates are available and there are 0 days of survival (i.e., date last contact= date of diagnosis)
- Complete dates are available and there are more than 0 days of survival (i.e. date last contact > date diagnosis)
- Incomplete dates are available and there could be zero days of follow-up (i.e., known components are equal, e.g. 99/99/2006 and 10/02/2006)
- Incomplete dates are available and there cannot be zero days of follow-up (i.e., any difference in known date components, e.g. 02/99/2006 and 03/99/2006)

**Seq\_Num (Sequence number):** This variable provides information about the sequence or order of multiple primary cancers. Sequence Number-Central describes the number and sequence of all reportable malignant, in situ, benign, and borderline

primary tumors, which occur over the lifetime of a patient. The sequence number may change over the lifetime of the patient. If an individual previously diagnosed with a single reportable malignant neoplasm is subsequently diagnosed with a second reportable malignant neoplasm, the sequence code for the first neoplasm changes from 00 to 01.

This sequence number counts all tumors that were reportable in the year they were diagnosed even if the tumors occurred before the registry existed, or before the registry participated in the SEER Program. The purpose of sequencing based on the patient's lifetime is to truly identify the patients for survival analysis who only had one malignant primary in their lifetimes.

**First\_Primary (First malignant primary indicator):** This variable was based on all the tumors in SEER. Tumors not reported to SEER were assumed malignant.

**Primary\_Rules (Primary by international rules):** This variable was created using IARC multiple primary rules. Did not include benign tumors or non-bladder in situ tumors in the algorithm. No tumor information was modified on any records.

**Follow\_Up\_Year (Year of Follow up):** This variable contains information about the year of follow-up data.

**Death\_Year (Year of death):** Year of death for patients who died before the study cutoff, as coded in the year of follow-up recode.

**Cancer\_Site (Detailed Malignancy Site):** This variable provides detailed information about the site of cancer cases, including malignant and in situ cases.

**Report\_Source (Type of Reporting Source):** The Type of Reporting Source identifies the source documents that provided the most complete information when

abstracting the case. This is not necessarily the original document that identified the case; rather, it is the source that provided the most complete information, for instance, Hospitals, radiation or oncology centers, physician's offices, nursing homes, autopsy/death certificates, and other hospital outpatients.

**Marital\_Status (Marital status at diagnosis):** The variable demonstrates the patient's marital status at the time of diagnosis of the malignancy. If the patient has multiple tumors, marital status may be different for each tumor. The sub-categories of the variable are as follows:

- Single (never married), Married (including common law), Separated, Divorced, Widowed, Unmarried or Domestic Partner (same sex or opposite sex, registered or unregistered, other than common law marriage) (effective for cases diagnosed 01/01/11 and forward), Unknown.

**Income (Median household income):** This variable contains information about the median household income adjusted for inflation to the year 2021.

**RUC (Rural-Urban Continuum Code):** Rural-Urban Continuum Codes were developed by the United States Department of Agriculture (USDA).

- Rural-urban continuum Codes form a classification scheme that distinguishes metropolitan (metro) counties by the population size of their metro area, and nonmetropolitan (nonmetro) counties by degree of urbanization and adjacency to a metro area or area.

### 3.4 TMN classification

An approach for categorizing malignancies is called the TNM Classification. It can help with prognostic cancer staging and is mostly applied to solid tumors. A standardized classification scheme facilitates improved information exchange and cross-population research while also enhancing communication between providers. As shown below, the system's foundation is the evaluation of the tumor, regional lymph nodes, and distant metastases.

- T stands for tumor. It is used to characterize the main tumor's size and the extent of its infiltration of nearby tissues. T0 denotes the absence of any tumor evidence, but T1-T4 was used to determine the tumor's size and extent, with T1-T4 showing gradual expansion and invasiveness. The evaluation of T-values varies according to the anatomic structures that are involved.
- N stands for nodes. It is used to indicate a tumor's localized infiltration of lymph nodes. Fluid from body tissues is absorbed into lymphatic capillaries and travels to the lymph nodes, where lymph nodes serve as biological filters. While N1-N3 show some degree of nodal spread with a gradually distal spread from N1 to N3, N0 indicates no regional nodal spread. The evaluation of N-values varies depending on the type of tumor and the lymph node drainage pattern in the area.
- M stands for Metastasis. It is utilized to determine whether the primary tumor's distant metastases are present. When a tumor spreads beyond regional lymph nodes, it is said to have metastasized. In cases where there

is no evidence of distant metastasis, a tumor is classed as MO, and in cases where there is, it is classified as M1.

### **Clinical Significance of TMN classification**

The TNM system aids in determining the disease's anatomical extent, and the three elements taken together can be used to determine the tumor's overall stage. This system simplifies cancer staging, with malignancies classified as I-IV, with IV being the most severe. Carcinoma in situ is referred to as stage 0, meaning that although it is not currently malignant, it may develop into cancer in the future. Stage Vis reserved for Wilms tumors and is the result of involvement in both kidneys at the time of initial diagnosis. The following is a condensed description of cancer staging and how it relates to the TNM classification.

- Stage 0 - Indicates carcinoma in situ. Tis, NO, MO.
- Stage I - Localized cancer. T1-T2, NO, MO.
- Stage II - Locally advanced cancer, early stages. T2-T4, NO, MO.
- Stage III - Locally advanced cancer, late stages. T1-T4, N1-N3, MO.
- Stage IV - Metastatic cancer. T1-T4, N1-N3, M1.

A higher mortality rate and more severe cancer are linked to progressive cancer staging.

### **3.5 Inclusion And Exclusion Criteria**

The SEER Registry collects data on T and N classification, age at diagnosis, sex, race, primary site, tumor grade, and tumor size. Using the SEER 1992-2020

database released April 2023. Patients with complete data on sociodemographic data, clinicopathological and histologic types included, and therapy alternatives were included. The variables included were Age Group, Gender, Year of Diagnosis, Race and Origin, Cancer Site, Diagnostic Confirmation, Clinical Grade, Laterality, Histology Behavior, Rare Tumors, Derived T Stage, Derived N Stage, Derived M Stage, Derived Stage Group, Scope of LN Surgery, Other Surgical Procedures, Surgery and Radiation Sequence, Radiation Type, Systemic Therapy and Surgery Sequence, Lymph Node Size, Bone Metastasis at Diagnosis, Brain Metastasis at Diagnosis, Liver Metastasis at Diagnosis, Lung Metastasis at Diagnosis, Distant Lymph Node Metastasis at Diagnosis, Other Distant Metastasis at Diagnosis, Cause-Specific Death Classification, Survival Months, Sequence Number, First Malignant Primary, Primary by International Rules, Year of Follow-up Recode, Year of Death Recode, Detailed Cancer Site, Reporting Source, Marital Status, Median Household Income, and Rural-Urban Continuum Code. However, about 70,000 patients with missing information, duplication or repetition of variables, and irrelevant variables were excluded.

### **3.6 Statistical Analysis**

#### ***3.6.a Data Cleaning***

All data were extracted from the SEER database through the NCI's SEER \* Stat software (Surveillance Research Program, National Cancer Institute SEER \* Stat software). All data cleaning was performed using R statistical software. (Chu, 2020) The patient cases with missing data, duplication or repetition of variables, and irrelevant variables were removed.

### ***3.6.b Descriptive Analysis***

Descriptive analysis was performed to summarize the features of the dataset. A univariate analysis including the calculation of the frequencies of each variable aided in independently exploring each variable in the dataset. Bivariate analysis namely the Pearson chi-square test was used to determine the prognostic features categorized in terms of sociodemographic, clinicopathological, histological, and treatment alternatives positively correlated with the outcome i.e. the year of death. Two-sided  $p < 0.05$  was considered statistically significant.

## **3.7 Machine Learning (ML)**

### ***3.7.a Definition of Machine Learning Models***

We used five Machine Learning models for outcome prediction-Lasso regression (LR), Random Forest (RF), extreme gradient boosting (XGBOOST), and k-nearest neighbors (KNN). (Peng, 2022)

- **Lasso Regression (LR):**

A linear regression method called Lasso Regression selects variables and applies regularization to enhance the statistical model's interpretability and prediction accuracy. It applies L1 regularization, which imposes a penalty equal to the amount of the coefficients' absolute values. As a result, certain coefficients become precisely zero, which essentially carries out variable selection. When working with high-dimensional data-where the number of features is high relative to the number of samples-Lasso Regression is especially helpful.

- **Random Forest (RF):**

Regression and classification are two applications for the ensemble learning technique known as Random Forest. During training, it creates a large number of decision trees, and then it outputs the class that is the mean prediction (regression) or the mode of the classes (classification) of the discrete trees. Random Forest enhances the accuracy and robustness of the model while decreasing overfitting, hence surpassing the performance of a single decision tree.

- **Extreme Gradient Boosting (XGBOOST):**

The optimized distributed gradient boosting library known as Extreme Gradient Boosting, or XGBoost, is made to be incredibly effective, adaptable, and portable. It is a fast and efficient implementation of gradient-boosted decision trees. XGBoost is well-known for its speed and effectiveness in machine learning competitions.

- **K-Nearest Neighbors (KNN):**

A straightforward instance-based learning technique for regression and classification is K- Nearest Neighbors. In KNN, an object is assigned to the class most frequent among its k nearest neighbors (k is a positive number, usually small), based on a majority vote of its neighbors.

### ***3.7.b Data pre-processing***

The data preprocessing procedure included centering, standardization/normalization of variables, and defining the outcome variable. It was done with recipe packages in R using Machine Shop Package.

### ***3.7.c Prediction modeling***

These Machine Learning models were built-in R using the Machine Shop Package and utilized in the prediction of oral cancer prognosis (Chu, 2020). In order to prevent overfitting or underfitting of the data, the model's training and validation were performed under 5-fold cross- validation.

Model assessment was conducted using Brier score, accuracy, specificity (true negative), and sensitivity (true positive), ROC AUC score. Integrated Brier score (IBS) was as the primary evaluation index to evaluate the prediction performance of the models.

## CHAPTER 4

### RESULTS

*Table 1. Sociodemographic Features*

Variable	Categories [codes]	Frequency	Percentage (%)
Age	05-09 years	5	0.061
	50-54 years	638	7.80
	55-59 years	975	11.92
	60-64 years	1201	14.69
	65-69 years	1224	14.97
	70-74 years	1170	14.30
	75-79 years	795	9.72
	80-84 years	604	7.38
	85+ years	532	6.50
	10-14 years	10	0.12
	15-19 years	23	0.28
	20-24 years	34	0.41
	25-29 years	80	0.97
	30-34 years	126	1.54
	35-39 years	153	1.87
	40-44 years	227	2.77
	45-49 years	380	4.64
Sex	Female	2991	36.58
	Male	5186	63.42
Year of Diagnosis	2018	2648	32.38
	2019	2721	33.28
	2020	2808	34.34

**Table 1. Sociodemographic Features (cont.)**

Race	Hispanic (All Races)	850	10.40
	"Non-Hispanic American Indian/Alaska Native	63	0.77
	Non-Hispanic Asian or Pacific Islander	898	10.98
	Non-Hispanic Black	395	4.83
	Non-Hispanic Unknown Race"	144	1.76
	Non-Hispanic White	5827	71.26
Marital Status (at Diagnosis)	Divorced	662	8.10
	Married (including common law)	4453	54.46
	Separated	59	0.72
	Single (never married)	1424	17.41
	Unknown	844	10.32
	Unmarried or Domestic Partner	81	0.99
	Widowed	654	8.00
Household Income	\$35,000 - \$39,999	31	0.38
	\$40,000 - \$44,999	20	0.24
	\$45,000 - \$49,999	155	1.90
	\$50,000 - \$54,999	173	2.12
	\$55,000 - \$59,999	370	4.52
	\$60,000 - \$64,999	312	3.82
	\$65,000 - \$69,999	513	6.27
	\$70,000 - \$74,999	932	11.40
	\$75,000+	5705	69.77
	< \$35,000	5	0.06
	Unknown/missing/no match/Not 1990-2021	1	0.01
Rural/Urban	Counties in metropolitan areas ge 1 million pop	4886	59.75
	Counties in metropolitan areas of 250,000 to 1 million pop	1830	22.38
	Counties in metropolitan areas of lt 250 thousand pop	493	6.03

**Table 1. Sociodemographic Features (cont.)**

	Nonmetropolitan counties adjacent to a metropolitan area	485	5.93
	Nonmetropolitan counties not adjacent to a metropolitan area	464	5.67
	Unknown/missing/no match/Not 1990-2021	1	0.01
	Unknown/missing/no match (Alaska or Hawaii - Entire State)	18	0.22

Table 1. shows the sociodemographic distribution of oral cancer patients; Age, Sex, Year of death, race, marital status at the time of diagnosis, household income, and Urban/ Rural location. The majority of the patients were in the age range of 65-69 years. Most of the patients were males (63%) and non-Hispanic whites (71%). 69.77% of the total individuals had a household income of 75,000 dollars or more.

**Table 2. Clinicopathological & Histological Features**

Variable	Categories [codes]	Frequency	Percentage
Cancer site	Floor of the Mouth	1378	16.85
	Gum and other mouth	4086	49.97
	Lip	658	8.05
	Salivary Gland	1644	20.11
	Tongue	411	5.03
Diagnostic Confirmation	Clinical diagnosis only	13	0.16
	Positive exfoliative cytology, no positive histology	152	1.86
	Positive histology	7974	97.52
	Positive microscopic confirm, method not specified	1	0.01
	Radiography without microscopic confirm	21	0.26
	Unknown	16	0.20

**Table 2. Clinicopathological & Histological Features (cont.)**

Clinical Grade Classification	1	1203	14.71
	2	1595	19.51
	3	367	4.49
	9	4379	53.55
	A	54	0.66
	B	199	2.43
	C	361	4.41
	D	19	0.23
Laterality	Bilateral, single primary	12	0.15
	Left - origin of primary	1359	16.62
	Not a paired site	5451	66.66
	Only one side - side unspecified	3	0.04
	Paired site, but no information concerning laterality	15	0.18
	Right - origin of primary	1337	16.35
ICD-O-3 Risto-Malignant Tumors			
	Neoplasm, malignant	30	0.37
	Carcinoma, NOS	107	1.31
	Large cell neuroendocrine carcinoma	1	0.01
	Small cell carcinoma, NOS	10	0.12
	Non-small cell carcinoma	1	0.01
	Verrucous carcinoma, NOS	106	1.30
	Papillary squamous cell carcinoma	34	0.42
	Squamous cell carcinoma, NOS	4591	56.15
	Squamous cell carcinoma, keratinizing, NOS	1164	14.24
	Squamous cell carcinoma, large cell, nonkeratinizing, NOS	369	4.51
	Squamous cell carcinoma, spindle cell	26	0.32
	Squamous cell carcinoma, adenoid	4	0.05
	Lymphoepithelial carcinoma	10	0.12
	Basaloid squamous cell carcinoma	65	0.79
	Squamous cell carcinoma, HPV-positive	195	2.38

**Table 2. Clinicopathological & Histological Features (cont.)**

	Squamous cell carcinoma, HPV-negative	24	0.29
	Basal cell carcinoma, NOS	45	0.55
	Adenocarcinoma, NOS	82	1.00
	Basal cell adenocarcinoma	37	0.45
	Adenoid cystic carcinoma	219	2.68
	Neuroendocrine carcinoma, NOS	2	0.02
	Oxyphilic adenocarcinoma	10	0.12
	Clear cell adenocarcinoma, NOS	9	0.11
	Sclerosing sweat duct carcinoma	1	0.01
	Sebaceous adenocarcinoma	2	0.02
	Mucoepidermoid carcinoma	457	5.59
	Mucinous adenocarcinoma	1	0.01
	Infiltrating duct carcinoma, NOS	121	1.48
	Secretory carcinoma of breast	42	0.51
	Polymorphous low grade adenocarcinoma	53	0.65
	Acinar cell carcinoma	178	2.18
	Adenosquamous carcinoma	12	0.15
	Epithelial-myoepithelial carcinoma	47	0.57
	Malignant melanoma, NOS	13	0.16
	Mucosa! lentiginous melanoma	5	0.06
	Spindle cell melanoma, NOS		0.01
	Carcinoma in pleomorphic adenoma	61	0.75
	Carcinosarcoma, NOS	3	0.04
	Malignant myoepithelioma	39	0.48
Rare Cancer Type	3.1 Epithelial tumor of major salivary glands	1365	16.69
	3.2 Salivary gland type tumor of head and neck	302	3.69
	37 Malignant Melanoma of Mucosaand Extracutaneous	19	0.23
	5.1 Squamous cell carcinoma with variants of oropharynx	2048	25.05
	5.2 Other epithelial tumors of oropharynx	1	0.01
	56.5 Neuroendocrine carcinoma of other sites	13	0.16
	6.1 Squamous cell carcinoma with variants of oral cavity	3755	45.92

**Table 2. Clinicopathological & Histological Features (cont.)**

	6.2 Squamous cell carcinoma with variants of lip	629	7.69
	6.3 Other epithelial tumors of oral cavity and lip	45	0.55
T- classification	TO	20	0.24
	T1	2596	31.75
	T2	2069	25.30
	T3	1083	13.24
	T4	146	1.79
	T4a	1076	13.16
	T4b	161	1.97
	Tis	2	0.02
	Tx	1024	12.52
N- classification	NO	4403	53.85
	N1	1351	16.52
	N2	262	3.20
	N2a	174	2.13
	N2b	402	4.92
	N2c	223	2.73
	N3	18	0.22
	N3a	51	0.62
	N3b	518	6.33
	Nx	775	9.48
M- classification	MO	7962	97.37
	M1	215	2.63
Stage group	0	1	0.01
	1	2711	33.15
	2	1462	17.88
	3	860	10.52
	4	133	1.63
	4A	1195	14.61
	4B	639	7.81
	4C	169	2.07
	88	19	0.23
	99	988	12.08
Bone metastases	No	8108	99.16
	Yes	69	0.84
Brain metastases	No	8170	99.91

**Table 2. Clinicopathological & Histological Features (cont.)**

	yes	7	0.09
Liver metastases	No	8144	99.60
	yes	33	0.40
Lung metastases	No	8065	98.63
	yes	112	1.37
Distant Lymph node metastases	None; no lymph node metastases	8127	99.39
	Yes; distant lymph node metastases	50	0.61
Other Metastases	None; no other metastases	8141	99.56
	Yes; distant mets in known site(s) other than bone, brain, liver, lung, dist LN	36	0.44
Sequence number of primary tumor	10th of 10 or more primaries	1	0.01
	1st of 2 or more primaries	392	4.79
	2nd of 2 or more primaries	1558	19.05
	3rd of 3 or more primaries	360	4.40
	4th of 4 or more primaries	108	1.32
	5th of 5 or more primaries	27	0.33
	6th of 6 or more primaries	13	0.16
	7th of 7 or more primaries	3	0.04
	8th of 8 or more primaries	2	0.02
	One primary only	5712	69.85
	Unknown sequence number - federally required in situ or malignant tumors		0.01
First primary malignancy indicator	No	1950	23.85
	yes	6227	76.15

Table 2. shows the frequency distribution of the clinicopathological and histological features which include; cancer site, diagnostic confirmation, clinical grade classification, laterality, cancer type, TMN classification, stage group, metastases extending to bone, brain, liver, lung, and lymph nodes, sequence number of the primary tumor and first primary

malignancy indicator. 17% of the oral cancer patients had carcinoma of the floor of the mouth and 20% - of the salivary gland. 8% of oral cancer patients had moderately differentiated primary tumors, 3% had poorly differentiated and 14% had well differentiated tumors. 56% of the patients had squamous cell carcinoma. 66% of the oral cancer patients did not have a paired site. Metastases of bone, brain, lung, and liver were absent in 99% of oral cancer patients. About 50% of the oral cancer patients had Stage I - Localized cancer (T1-T2, NO, MO).

**Table 3. Treatment Options of Head & Neck Cancer Patients**

Variable	Categories	Frequency	Percentage (%)
Lymph Node Surgery	1 to 3 regional lymph nodes removed	384	4.70
	4 or more regional lymph nodes removed	3156	38.60
	Biopsy or aspiration of regional lymph node, NOS	576	7.04
	None	3915	47.88
	Number of regional lymph nodes removed unknown	32	0.39
	Sentinel lymph node biopsy	29	0.35
	Sentinel node biopsy and lymph node removed at different times	10	0.12
	Sentinel node biopsy and lymph node removed same/unstated time	8	0.10
	Unknown or not applicable	67	0.82
	Other Surgical Procedure related to regional/ distant metastases	Any combo of surgery procedure to other regional, distant lymph node, and/or distant site	3
	Non-primary surgical procedure performed	53	0.65
	Non-primary surgical procedure to distant lymph node(s)	6	0.07
	Non-primary surgical procedure to distant site	22	0.27
	Non-primary surgical procedure to other regional sites	121	1.48

**Table 3. Treatment Options of Head & Neck Cancer Patients (cont.)**

	None; diagnosed at autopsy	7943	97.14
	Unknown; death certificate only	29	0.35
Surgery and Radiation Therapy	Intraoperative radiation	1	0.01
	No radiation/ cancer-directed surgery	5356	65.50
	Radiation after surgery	2783	34.03
	Radiation before and after surgery	12	0.15
	Radiation prior to surgery	24	0.29
	Surgery both before and after radiation	1	0.01
Radiation Therapy	Beam Radiation	3886	47.52
	Combination of beam with implants or isotopes	4	0.05
	None/unknown	3936	48.14
	Radiation, NOS method or source not specified	8	0.10
	Radioactive implants ( includes brachytherapy )	4	0.05
	Recommended, unknown if administered	69	0.84
	Refused	270	3.30
Sequence of Systemic Therapy or Surgical Procedure	No systemic therapy and/or surgical procedures	6750	82.55
	Sequence unknown	1	0.01
	Surgery both before and after systemic therapy	5	0.06
	Systemic therapy after surgery	1351	16.52
	Systemic therapy before surgery	40	0.49
	Systemic therapy both before and after surgery	30	0.37

Table 3. includes the distribution of oral cancer patients based on treatment alternatives like lymph node surgery, radiation therapy, sequence of systemic therapy and surgical procedure, and other surgical procedures related to regional/ distant metastases. 38% of the oral cancer patients underwent lymph node surgery where 4 or more regional lymph nodes were

removed. 16% of the oral cancer patients took systemic therapy after surgery. 47% of oral cancer patients received beam radiation therapy. 34% of the oral cancer patients underwent surgery followed by radiation therapy.

**Table 4. Association between Oral Cancer-Related Mortality and Predictors of Oral Cancer mortality**

Variable	Sub-categories	Died due to oral cancer (in %)	Alive or died due to other causes(in %)	P-value
Sex	Male	61	64	0.09
	Female	39	36	
Age	40-44 years	3	7	<0.01
	45-49 years	4	6	
	50-54 years	5	8	
	55-59 years	9	12	
	60-64 years	14	15	
	65-69 years	9	16	
	70-74 years	14	14	
	75-79 years	11	10	
	80-84 years	11	7	
	85+ years	20	5	
Year of Diagnosis	2018	50	31	< 0.05
	2019	37	33	
	2020	13	36	
Race	1	11	10	0.02
	2	1	1	
	3	13	11	
	4	5	5	

	5	0	2	
	6	70	71	
Cancer site	1	14	17	< 0.01
	2	55	49	
	3	2	9	
	4	22	20	
	5	8	5	
Diagnostic Confirmation	1	0	0	< 0.01
	2	2	2	
	3	94	98	
	4	0	0	
	5	2	0	
	6	1	0	
Race	1	11	10	< 0.01
	2	1	1	
	3	13	11	
	4	5	5	
	5	0	2	
	6	70	71	
Grade	1	16	15	< 0.01
	2	30	18	
	3	11	4	
	9	36	55	
	A	0	1	
	B	2	3	
	C	5	4	
	D	0	0	
Laterality	1	0	0	0.02
	2	16	17	
	3	71	66	
	4	0	0	
	5	0	0	
	6	13	17	
Rare Cancer Type	1	14	17	< 0.01
	2	1	4	
	3	0	0	
	4	21	25	
	5	0	0	

	6	0	0	
	7	61	44	
	8	2	8	
	9	0	0	
T stage	T0	0	0	< 0.01
	Tis	0	0	
	TX	16	12	
N stage	N0	30	56	< 0.01
	N1	15	17	
	N2	2	3	
	N2a	4	2	
	N2b	10	4	
	N2c	8	2	
	N3	0	0	
	N3a	0	0	
	N3b	19	5	
	Nx	11	9	
M stage				
	M1	11	2	< 0.01
	M2	89	98	
Stage Group Classification	0	7	0	< 0.01
	1	0	36	
	2	9	19	
	3	10	11	
	4	3	1	
	4A	26	13	
	4B	22	6	
	4C	9	1	
	88	0	0	
	99	13	12	
Lymph Node Surgery	1	2	5	< 0.01
	2	38	39	
	3	5	7	
	4	54	47	
	5	1	0	
	6	0	0	
	7	0	0	
	8	0	0	
	9	0	2	

Surgical Procedures	1	0	0	0.55
	2	0	1	
	3	0	0	
	4	0	0	
	5	2	2	
	6	98	97	
	7	0	0	
Sequence of Surgery and Radiation	1	0	1	0.16
	2	70	65	
	3	30	34	
	4	0	0	
	5	0	0	
	6	0	0	
Radiation Therapy	1	50	47	< 0.01
	2	0	0	
	3	40	49	
	4	0	0	
	5	0	0	
	6	1	0	
	7	9	4	
Sequence of systemic therapy and surgical procedure	0	80	83	< 0.01
	1	0	0	
	2	0	0	
	3	19	16	
	4	1	1	
	5	0	0	
Bone metastases	YES	4	1	< 0.01
	NO	96	99	
Brain metastases	YES	0	0	< 0.01
	NO	100	100	
Liver metastases	YES	2	0	< 0.01
	NO	98	100	
Lung metastases	YES	6	1	< 0.01
	NO	94	99	

Distant Lymph Node metastases	YES	3	0	< 0.01
	NO	97	100	
Other metastases	YES	3	0	< 0.01
	NO	97	100	
Conditions for calculating survival months	0	0	1	< 0.01
	1	97	98	
	2	3	1	
	3	0	0	
Sequence number of primary tumours	0	5	5	0.83
	1	1	1	
	2	19	19	
	3	5	4	
	4	1	1	
	5	0	0	
	6	0	0	
	7	0	0	
	8	0	0	
	9	69	70	
First primary malignancy indicator	YES	75	76	0.44
	NO	25	24	
IARC multiple primary rules	YES	94	96	0.10
	NO	6	4	
Death Year	Alive at last contact	0	100	< 0.01
	2020	50	0	
	2019	36	0	
	2018	15	0	
Cancer Site	1	8	5	< 0.01
	2	22	20	
	3	2	9	
	4	14	17	
	5	55	49	

Type of Reporting Source	1	93	91	< 0.01
	2	3	5	
	3	0	0	
	4	1	1	
	5	2	2	
	6	1	1	
Marital Status ( at the time of diagnosis )	0	11	8	< 0.01
	1	40	56	
	2	1	1	
	3	23	17	
	4	10	10	
	5	1	1	
	6	14	7	
Household Income	0	0	0	< 0.01
	1	0	0	
	2	1	0	
	3	1	0	
	4	2	1	
	5	5	5	
	6	3	4	
	7	6	6	
	8	17	11	
	9	64	70	
	10	0	0	
Rural-Urban Continuum Code	0	59	60	0.69
	1	21	22	
	2	6	6	
	3	6	6	
	4	7	6	
	5	0	0	
	6	0	0	

Table 4. shows the relationship between oral cancer predictors and oral cancer-related mortality. Death is a binary variable ( outcome variable ) denoted as Dead due to oral cancer which were 732 participants(9%); and Alive or dead of other causes were 7444 participants(91%). Out of 732, 61% of the males died from oral cancer. It was noted that 20% of the participants were over 85 years old. According to our model, a 5% significance level was taken into consideration as it is an industry-standard value. Based on the model's statistical summary, the Pearson Chi-square test (bivariate analysis) demonstrated that out of 38 variables, 29 were statistically significant as the p-value was less than 0.05. The 29 variables that were statistically significant are as follows- age, year of diagnosis, race, cancer site, histology, cancer grade classification, laterality of the cancer tumor, T-classification, N-classification, M-classification, stage group classification, lymph node surgery, radiation therapy, systematic therapy and surgery sequence, death year, cancer site classification, bone metastasis, brain metastasis, lung metastasis, liver metastasis, other distant metastasis, distant lymph nodes, Conditions for calculating survival months, sequence number of primary tumors, First primary malignancy indicator, IARC multiple primary rules, report source, marital status, and income.

## **PREDICTIVE MODELLING FOR ORAL CANCER RELATED MORTALITY**

### **A. Xgboost:**

XGBoost model showed a Brier Score of 0.0677, an accuracy of 91%, a 13% kappa statistic, an ROC AUC of 84%, a sensitivity of 99%, and less than 1% specificity.

Our variable importance table (Table 5) and plot (Figure 1) showed that out of 38 variables assessed in the model, 17 were found to be the most important predictors of OCRM. The most important predictors of OCRM (in descending order) were cancer stage group, age, T stage, Lymph node surgery, cancer site, tumor rarity, N stage, marital status, radiation, income, grade, lymph node size, surgery radiation sequence, race, histology, the sequence number of multiple primary cancers, side of a paired organ which tumor originated.

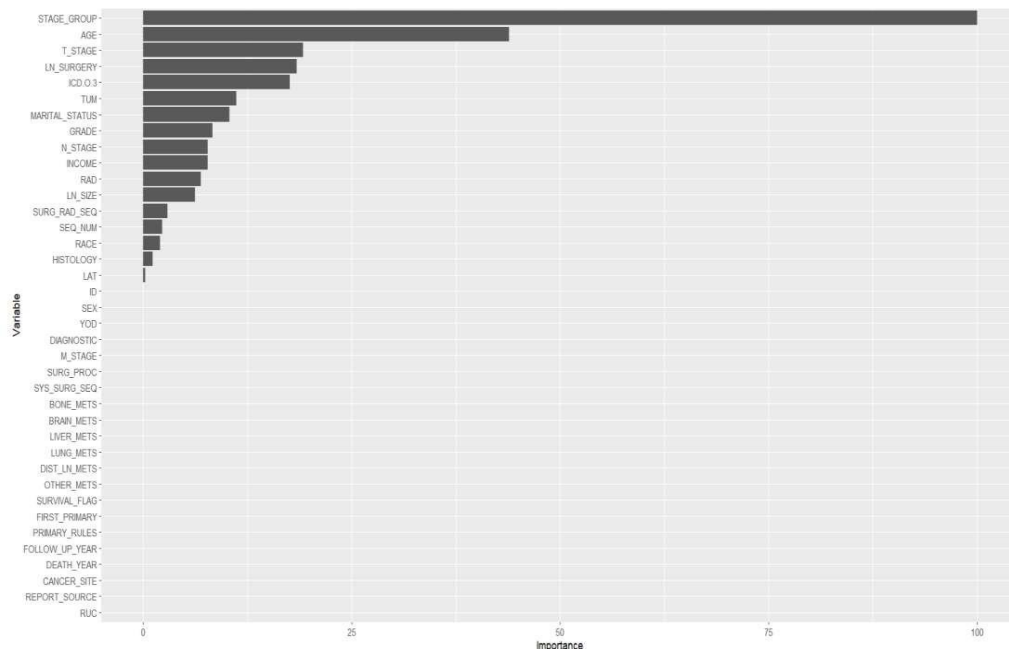


Figure 1. Variable Importance Plot - Xgboost

As shown in Figure2, the Xgboost machine learning model was fairly well calibrated at higher probabilities while the lower probabilities had somewhat poor calibration. Overall, the model was well calibrated which means that the observed mean values track well with the predicted mean values.

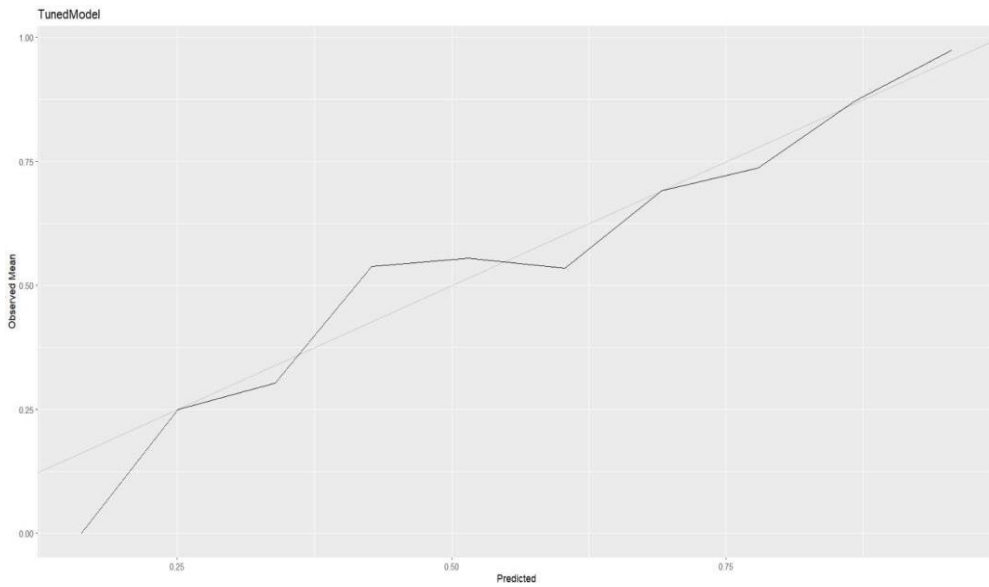


Figure 2. Calibration Plot - Xgboost

**Table 5. Variable Importance Table - Xgboost**

Variable	Permute Mean Brier
Stage group	100
Age	43.86
T stage	19.17
Lymph node surgery	18.43
ICD.O.3 classification	17.61
Tumor	11.14
Marital Status	10.37
Grade	8.31

**Table 5. Variable Importance Table – Xgboost (cont.)**

N stage	7.74
Income	7.73
Radiation	6.91
Lymph node size	6.24
Sequence of surgery and radiation therapy	2.90
Sequence Number of primary tumor	2.25
Race	2.02
Histology	1.12
Laterality	0.23

A. Lasso Regression:

Lasso regression model showed that the prediction performance of the Lasso Regression ML model showed a Brier Score of 0.0722, an accuracy of 91%, a 2% kappa statistic, an ROC AUC of 80%, a sensitivity of 99%, and less than 1% specificity.

Our variable importance table (Table 6) and plot (Figure 3) showed that out of 38 variables assessed in the model, 15 were found to be the most important predictors of OCRM. The most important predictors of OCRM (in descending order) were cancer stage group, age, T stage, grade, N stage, radiation, marital status, tumor rarity, lymph node surgery, ICD.O.3, cancer site, income, surgery radiation sequence, lymph node size, and surgery sequence.

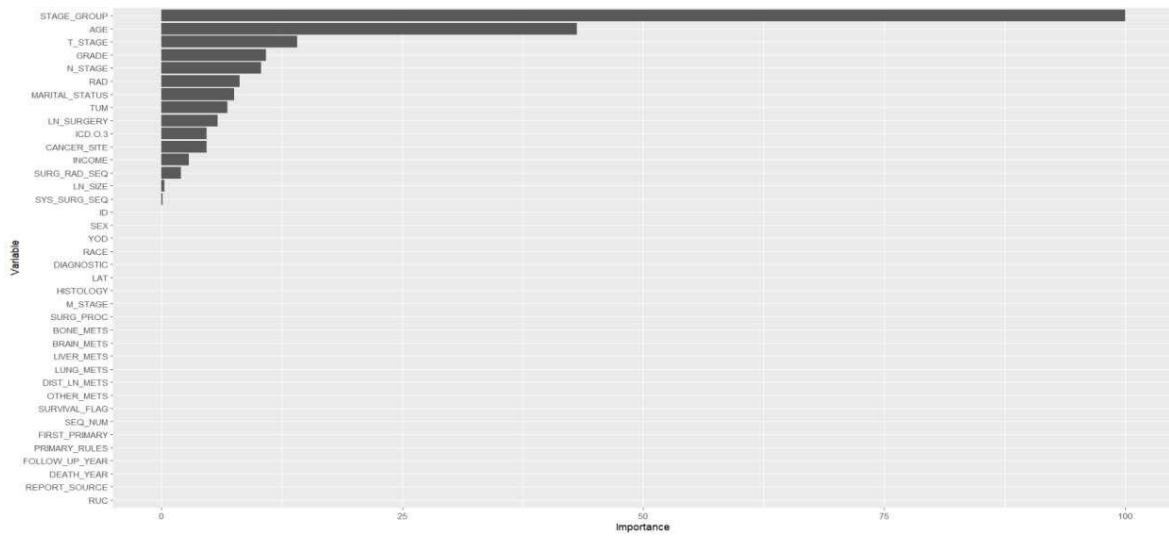


Figure 3. Variable Importance Plot - Lasso Regression

As shown in Figure 4. the calibration plot of the Lasso Regression machine learning model where it was observed that the higher probabilities were well calibrated while the lower probabilities did not do well. Although, overall, the plot was well calibrated.

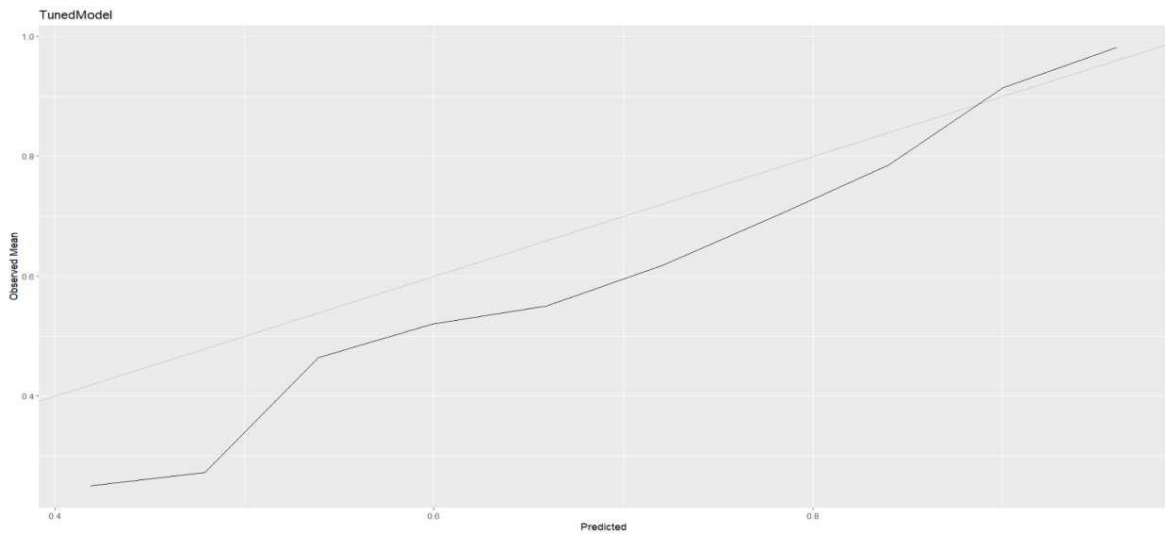


Figure 4. Calibration Plot - Lasso Regression

**Table 6. Variable Importance Table - Lasso Regression**

Variable	Permute mean brier
Stage Group	100
Age	43.13
T stage	14.11
Grade	10.82
N stage	10.32
Radiation	8.12
Marital Status	7.53
Tumor	6.85
Lymph node surgery	5.84
ICD.O.3 cancer classification	4.69
Cancer site	4.67
Income	2.86
Surgery Radiation Sequence	2.01
Lymph node size	0.30
Systemic surgery sequence	0.14

**B. Random Forest:**

Random Forest ML model showed the prediction performance of a Brier Score of 0.0725, an accuracy of 91%, a 14% kappa statistic, an ROC AUC of 81%, a sensitivity of 99%, and less than 1% specificity.

Our variable importance table (Table 7) and plot (Figure 5) showed that out of 38 variables assessed in the model, 21 were found to be the most important predictors of OCRM. The most important predictors of OCRM (in descending order) were cancer stage group, age, lymph node size, T-stage, grade, marital status, income, lymph node surgery, N-stage, tumor rarity, histology, radiation therapy, sequence of radiation and surgery, ICD.O.3, the sequence number of multiple primary cancers, cancer site, race, side of a paired organ which tumor originated from, sex, and surgery sequence.

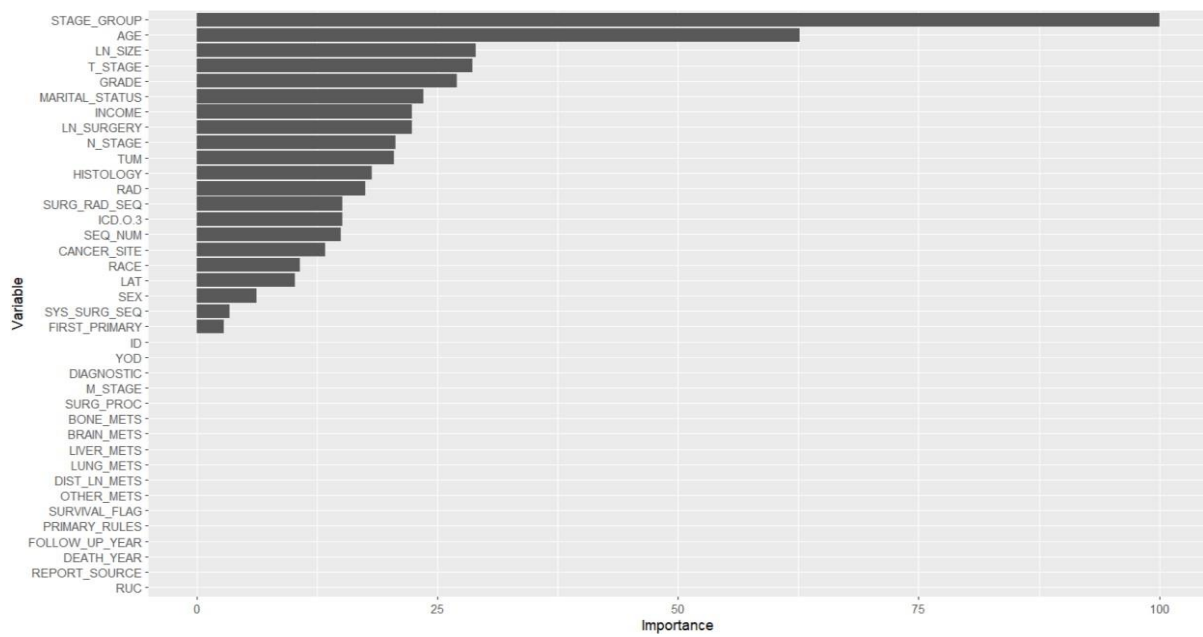


Figure 5. Variable Importance Plot - Random Forest

As shown in Figure 6, the calibration plot of Random Forest machine learning model where it was noted that the plot wasn't well calibrated. The plot could not track well the observed mean values with the predicted mean values.

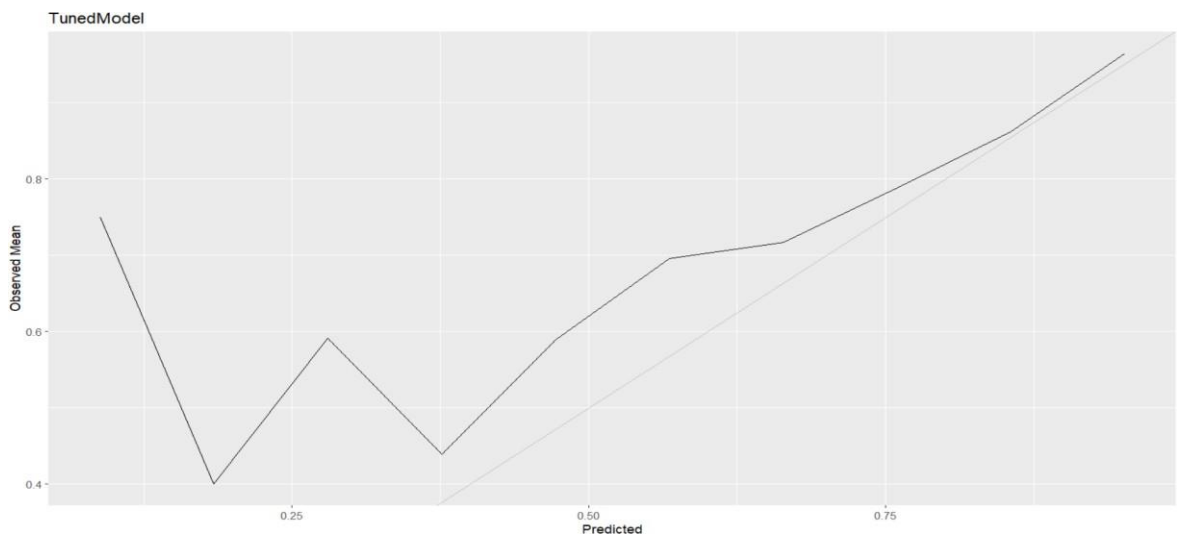


Figure 6. Calibration Plot - Random Forest

**Table 7. Variable Importance Table - Random Forest**

Variable	Permute mean brier
Stage Group	100
Age	62.66
Lymph node size	28.99
T stage	28.59
Grade	26.97
Marital Status	23.49
Income	22.37
Lymph Node Surgery	22.30
N stage	20.65
Tumor	20.46
Histology	18.13
Radiation	17.47
Surgery Radiation Sequence	15.09
ICD.0.3 Classification	15.07
Sequence number of Primary Tumor	14.94
Cancer Site	13.30
Race	10.71
Laterality	10.16
Sex	6.18
Systemic Surgery Sequence	3.38

C. K- Nearest Neighbors (KNN)-

K- Nearest Neighbors ML model showed prediction performance of a Brier Score of 0.0785, an accuracy of 91%, a 5% kappa statistic, an ROC AUC of 72%, a sensitivity of 99%, and less than 1% specificity.

Our variable importance table (Table 8) and plot (Figure 7) showed that out of 38 variables assessed in the model, 20 were found to be the most important predictors of OCRM. The most important predictors of OCRM (in descending order) were lymph node size, histology, age, cancer stage group, T-stage, N-stage, sequence number of multiple primary cancers, income, lymph node surgery, tumor rarity, marital status, radiation, race, cancer grade, systemic surgery sequence,

sequence of radiation and surgery, side of a paired organ which tumor originated from, ICD.0.3, cancer site, and sex.

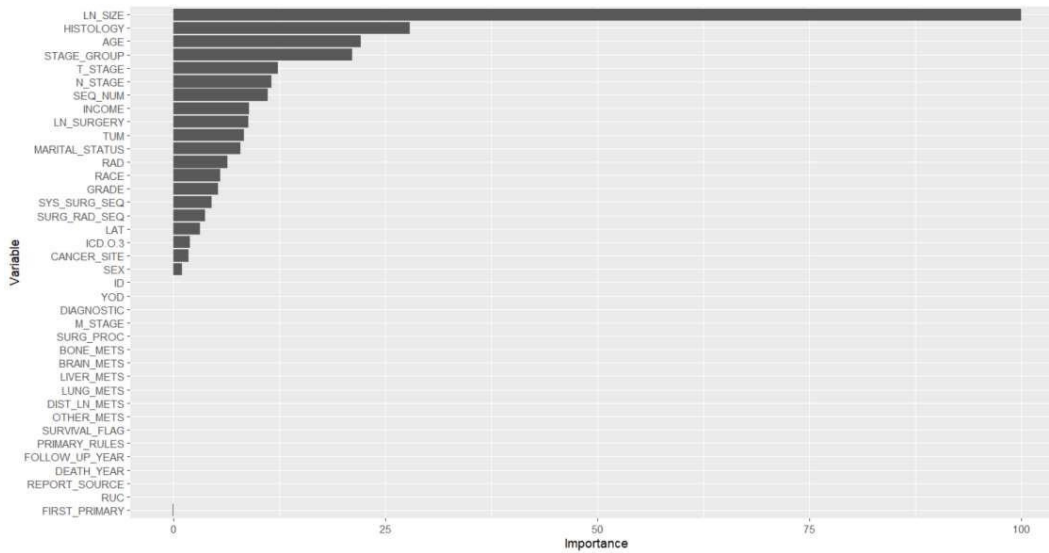


Figure 7. Variable Importance Plot- K- Nearest Neighbors

Figure 8. shows the calibration plot of K- Nearest Neighbor machine learning model where it was noted that the plot could not track well the observed mean values with the predicted mean values. Thus, the plot was not well calibrated.

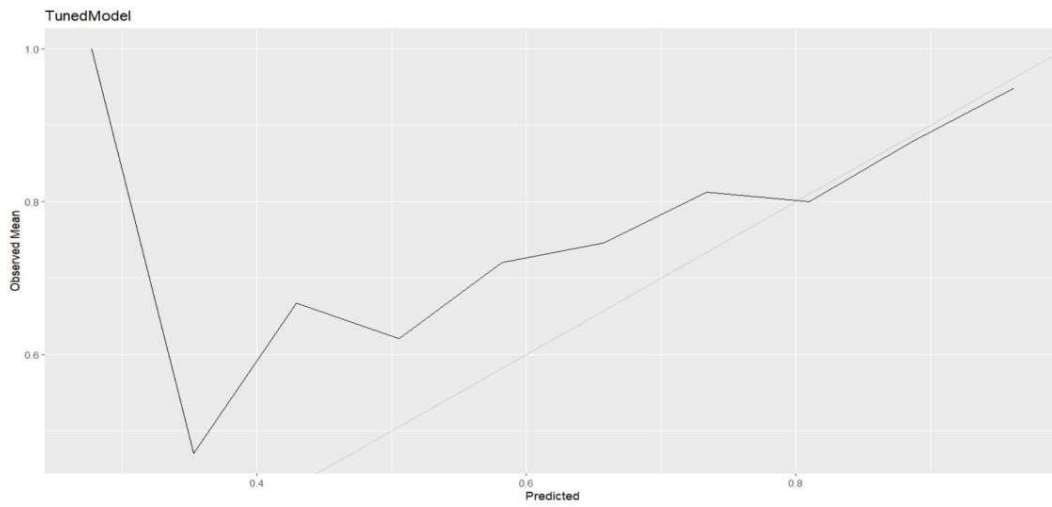


Figure 8. Calibration Plot- K- Nearest Neighbors

**Table 8. Variable Importance Table - K- Nearest Neighbors**

Variable	Permute mean brier
Lymph node size	100
Histology	28
Age	22.06
Stage Group	21.08
T stage	12.28
N stage	11.51
Sequence number of primary tumor	11.11
Income	8.91
Lymph node surgery	8.85
Tumor	8.32
Marital Status	7.88
Radiation	6.34
Race	5.51
Grade	5.24
Systemic Surgery Sequence	4.45
Surgery and Radiation sequence	3.68
Laterality	3.11
ICD.O.3 classification	1.95
Cancer site	1.79
Sex	0.96

**Table 9. Comparison Of Machine Learning Models - extreme gradient boosting machine versus lasso regression versus random forest versus K-Nearest Neighbour**

Statistics/prediction model	XGBT	Lasso	RF	KNN
Brier Score	0.07	0.07	0.07	0.08
Accuracy	91.31%	91.12%	90.81%	90.88%
Kappa Statistics	13%	2%	14%	5%
ROCAUC	83.49%	80.78%	81.10%	71.93%
Sensitivity	99.42%	99.94%	98.69%	99.46%
Specificity	8.81%	1.39%	10.71%	3.69%

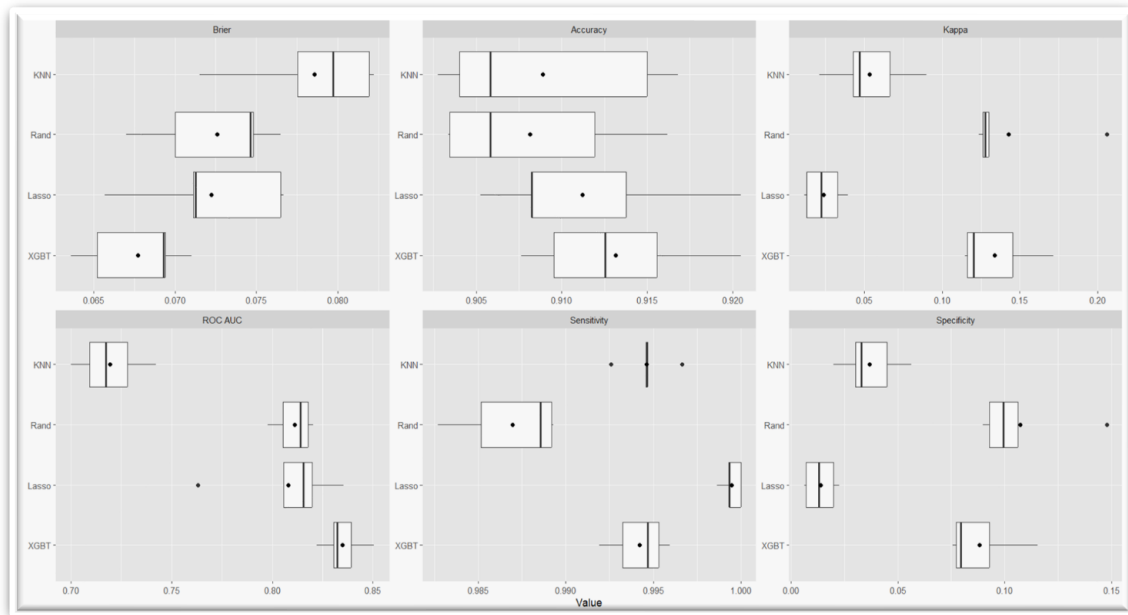


Figure 9. Boxplot representation of performance comparison and selection of Machine Learning Models

Table 9. shows a comparison of machine-learning models. XGBT had the highest ROC AUC (83%) whereas the KNN model had the lowest ROC AUC of 72%. The sensitivity for all Machine Learning models was about 99%. In comparison with other machine learning models, XGboost and Random Forest machine learning models had the

highest specificity of 8.81% and 10.71% respectively. Furthermore, the Brier score was highest with the KNN model and lowest with XGBT. Hence, across all models, the XGBT model was the best-performing machine learning model. The comparison has been visualized in Fig. 9 with the boxplot representation of the performance comparison of machine learning models. It was based on the Brier score, Accuracy, Cohen's Kappa statistic, ROC AUC, Sensitivity, and Specificity showing that the XGBT model has the highest accuracy and ROC AUC and lowest Brier score.

## CHAPTER 5

### DISCUSSION

Head and neck cancer (HNC) is the sixth most frequent cancer in the world, accounting for more than 500,000 new cases each year (Lin, 2021). It comprises extremely distinct tumors produced from stratified epithelial cells of numerous anatomical subsites, primarily the oral cavity, nasal cavity, larynx, and throat, with squamous cell carcinoma being the most prevalent histological type (Lin, 2021). Doctors prefer to forecast patients' survival based on their tumor status, cancer stages, and ages, among other factors; nevertheless, it is difficult to define patients' clinicodemographic features more accurately. As a result, it is critical to develop accurate prognostic prediction models based on numerous clinicopathologic criteria to help doctors make personalized clinical decisions and select appropriate treatment options (Peng, 2022).

In oncology research, machine learning algorithms that automate the creation of analytical models have been used for the past ten years in an effort to improve prediction and make more accurate clinical outcome forecasts. Their widespread appeal stems from their alleged capacity to progressively identify patterns, gather data, and go through automated training based on data input-especially complex non-homogenous data-resulting in the production of clinical forecasts with little assistance from humans. While some algorithms-support vector machines, boosted decision trees, decision forests, and artificial neural networks, in particular-have been shown to have some predictive accuracy, before machine learning is widely applied to clinical practice, it is necessary to validate its predictive power by

examining the course of the disease in well-defined OSCC patient cohorts. (Chu, 2020).

The sociodemographic distribution of oral cancer patients in our study included age, sex, year of death, race, marital status at the time of diagnosis, household income, and urban/ rural location. The majority of the patients were older adults. Most of the patients were males and non- Hispanic whites. A greater number of individuals had a household income of around \$75,000+. The distribution of the clinicopathological and histological features include cancer site, histology confirmation, clinical grade classification, laterality of primary tumor, cancer type, TMN classification, stage group, metastases extending to bone, brain, liver, lung, and lymph nodes, the sequence number of the primary tumor and first primary malignancy indicator. Among all oral cancer patients, the most common sites were the floor of the mouth, salivary gland, and gums.

Most of the oral cancer patients had well-differentiated tumors followed by moderately differentiated primary tumors, and then, poorly differentiated tumors. About half of the oral cancer patients had Stage I - Localized cancer. Metastases of bone, brain, lung, and liver were absent in almost all oral cancer patients. The treatment alternatives assessed in our study included lymph node surgery, radiation therapy, sequence of systemic therapy and surgical procedures, and other surgical procedures related to regional/ distant metastases. Our study showed that the majority of the oral cancer patients received beam radiation therapy followed by lymph node surgery where 4 or more regional lymph nodes were removed and underwent radiation therapy after surgery. The least number of oral cancer patients took systemic therapy after surgery.

The xgboost and lasso regression machine learning models were fairly well calibrated at higher probabilities while the lower probabilities had somewhat poor calibration. Overall, the plots were well-calibrated. However, k- nearest neighbors and random forest were not well calibrated which means that the observed mean values tracked well with the predicted mean values.

A comparison of our predictive models showed XGBT to be the best-performing model when compared to the KNN, RF, and Lasso based on their Brier score and ROC AUC values.

XGBT had the highest ROC AUC followed by RF and then Lasso. The KNN model had the lowest ROC AUC. The XGBT model had the lowest Brier score which was the same as the Lasso and RF models while the KNN model had the highest Brier score. The sensitivity for all Machine Learning models was about 99%. In comparison with other machine learning models, XGboost and Random Forest machine learning models had the highest specificity. Hence, across all models, the XGBT model was the best-performing machine learning model to predict oral cancer-related mortality as it achieved the highest accuracy and ROC AUC with the lowest Brier score. Compared to Chu et al. (2020) study, our study utilized fewer machine learning models and performed fairly better in terms of accuracy performance. Our study found a higher accuracy in predicting oral cancer-related mortality and determining the predictors of oral cancer compared to the findings of Chu et al's (2020) study which had lower accuracy. Peng et al. (2022) also employed slightly more machine learning models with overall lower performance compared to our study in terms of ROC AUC. Our study also compared specificity for predicting oral cancer-related mortality which was relatively low compared to that of Chu et al

which had a higher specificity. The differences in performance might be due to differences in outcome measures, the type of model used, tuning parameters, and the type of dataset. For example, while our study found XGBT to be the best-performing model, Chu et al. (2020) and Peng et al. (2022) found decision tree (DT) and Random Forest models to be their best-performing models, respectively. Our study utilized the maximum number of prognostic features in comparison with Chu et al. (2020) which employed 34 prognostic features and Peng et al. (2022) which employed 25 prognostic features. Our study utilized a more expansive and comprehensive version of variables wherein we utilized 37 prognostic features. In our study, the most important predictors of OCRM (in descending order) were cancer stage group, age, T stage, Lymph node surgery, cancer site, tumor rarity, N stage, marital status, radiation, income, grade, lymph node size, surgery radiation sequence, race, histology, the sequence number of multiple primary cancers, side of a paired organ which tumor originated. This is comparable to Peng et al that found age, and tumor size to have the strongest positive effects on the Overall Survival and Cancer-Specific Survival prediction of HNC with several tumor sites.

In summary, Our study determined the effectiveness of machine learning in Oral Cancer research and assessed the most important predictors of OCRM. Although our study had lower percentages of specificity, it highlighted higher percentages of accuracy, ROC AUC, and sensitivity in detecting oral cancer-related mortality.

### **5.1 Strength and Weakness**

This study is one of the first to create an algorithm for predicting oral cancer

mortality in the SEER database, paving the way for future research in several medical and dental domains. Employing the advancements of machine learning and incorporating a large sample size of the SEER database, our study developed four machine-learning models to predict oral cancer-related mortality among adults in oral cancer patients integrating sociodemographic and clinicopathological features. This study had several drawbacks. Firstly, there were certain restrictions on the SEER database data. For instance, there were insufficient details about immunotherapy, radiation dosage, and lifestyle choices including drinking and smoking. Machine learning algorithms rely on the accuracy and quality of the data that is entered. Additionally, due to the lack of genetic profiling, biomarker analyses, and advanced histopathological imaging in conventional medical records the predictive analysis for determining oral cancer-related mortality can be challenging. Our study presented with low specificity across all machine learning models. Thus, it is more likely to produce a high number of false positives and may incorrectly identify disease or illness in individuals when it is not present. In order to address this drawback in the future, we can define some useful metrics, such as precision and recall. Additionally, we can combine them into a new more powerful metric called the F-score. Data scientists widely use these metrics to evaluate their models.

## **5.2 Future Directions**

### ***5.2.a Development of a Personalized Diagnosis App for Dental Health:***

Our primary goal is to create a user-friendly app that allows patients to securely upload their medical and dental history along with clinical images of any pathologies they may have. This comprehensive data collection will enable personalized diagnosis and treatment recommendations. Moreover, by integrating

virtual assistance from clinicians within the app, patients can receive expert guidance remotely, thereby breaking down barriers to accessing dental care, especially in areas where traditional in-person visits are challenging. This initiative not only enhances convenience for patients but also serves as a crucial step forward in the realm of teledentistry, fostering better communication and collaboration between patients and dental professionals.

***5.2.b Integration of Genomic Data for Enhanced Cancer Analysis and Risk Assessment:***

Our goal is to transform cancer analysis in dentistry by merging genomic datasets with our current (SEER) database. Through this integration, we will be able to determine the genetic markers linked to particular tumors, facilitating more precise risk assessment and individualized treatment plans. By using sophisticated analytics, we can forecast each patient's reaction to different treatment modalities, improving patient outcomes. This novel method demonstrates our dedication to providing individualized dental interventions that put the needs of our patients first, while also deepening our understanding of the genetic foundations of oral malignancies.

***5.2.c Implementation of Large Language Model (LLM) for Clinician Decision Support:***

A major step forward in using artificial intelligence to improve decision-making processes is the incorporation of the Large Language Model (LLM) into our clinical workflow. LLM simplifies discussions and makes it easier to create customized treatment regimens for each patient by giving clinicians text-based and voice-free inputs. By instantly combining a plethora of data, scientific literature, and patient information, this advanced technology enhances the ability of clinicians and helps

them make more educated and effective clinical judgments. By utilizing LLM, we enable dental professionals to provide patients with optimal care that is customized to their personalized requirements and preferences, raising the bar for dental care and improving the overall satisfaction of patients.

## **CHAPTER 6**

### **CONCLUSION**

In this paper, we reviewed the foundations of machine learning and looked at their application to cancer prognosis and prediction. The predictive values of different types of oral cancer were examined using four Machine Learning Models. Based on the results, it was determined that the XGBoost model performed the best. This suggests that the XGBoost machine learning model could be used for clinical decision-making and patient consultation for a variety of patients with oral cancer. To summarize, our machine learning model demonstrated good prediction performance in estimating the death rate from oral cancer using clinicopathological and histological factors sourced from the national cancer registry. Additionally, the XGBoost machine learning model's mortality prediction was significantly influenced by age and the cancer stage group.

## REFERENCES

1. K Alabi, R. O., Elmusrati, M., Leivo, I., Almangush, A., & Makitie, A. A. (2023). Advanced-stage tongue squamous cell carcinoma: A machine learning model for risk stratification and treatment planning. *Acta Otolaryngologica*, 143(3), 206-214. <https://doi.org/10.1080/00016489.2023.2172208>
2. Fatapour, Y., Abiri, A., Kuan, E. C., & Brody, J. P. (2023). Development of a machine learning model to predict recurrence of oral tongue squamous cell carcinoma. *Cancers*, 15(10), 2769. <https://doi.org/10.3390/cancers15102769>
3. Peng, J., Lu, Y., Chen, L., Qiu, K., Chen, F., Liu, J., Xu, W., Zhang, W., Zhao, Y., Yu, Z., & Ren, J. (2022). The prognostic value of machine learning techniques versus Cox regression model for head and neck cancer. *Methods*, 123-132.
4. K Alabi, R. O., Makitie, A. A., Pirinen, M., Elmusratia, M., Leivo, I., & Almangush, A. (2021). Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer. *International Journal of Medical Informatics*, 104313.
5. Du, M., Haag, D. G., Lynch, J. W., & Mittinty, M. N. (2020). Comparison of the tree-based machine learning algorithms to Cox regression in predicting the survival of oral and pharyngeal cancers: Analyses based on SEER database. *Cancers*.
6. Tewari, P., Kashdan, E., Walsh, C., Martin, C. M., Parnell, A. C., & O'Leary, J. J. (2021). Estimating the conditional probability of developing human papillomavirus-related oropharyngeal cancer by combining machine learning and inverse Bayesian modelling. *PLoS Computational Biology*, 17(8), e1009289. <https://doi.org/10.1371/journal.pcbi.1009289>
7. Wong, T. S. C., & Wiesenfeld, D. (2018). Oral cancer. *Australian Dental Journal*, 63 (1 Suppl), S91- S99. <https://doi.org/10.1111/adj.12594>
8. Barr, A., Feigenbaum, E. A., & Cohen, P.R. (1981). *The Handbook of Artificial Intelligence* (Vols. 1- 3). William Kaufmann Inc.
9. Schwendicke, F., Samek, W., & Krois, J. (2020). Artificial intelligence in dentistry: Chances and challenges. *Journal of Dental Research*, 99(7), 769-774.

10. Ahmed, N., Abbasi, M. S., Halim, M. S. B., Zuberi, F., Qamar, W., Maqsood, A., & Alam, M. K. (2020). Artificial intelligence techniques: Analysis, application, and outcome in dentistry-A systematic review. *BioMed Research International*.
11. World Health Organization. (2020). Oral cancer. Retrieved from <https://www.who.int/cancer/prevention/diagnosis-screening/oral-cancer/en/>
12. Alhazmi, A., Alhazmi, Y., Makrami, A., et al. (2021). Application of artificial intelligence and machine learning for prediction of oral cancer risk. *Journal of Oral Pathology & Medicine*, 50, 444--450. <https://doi.org/10.1111/jop.13157>
13. Alabi, R. O., Elmusrati, M., Leivo, I., Almangush, A., & Makitie, A. A. (2023). Advanced-stage tongue squamous cell carcinoma: A machine learning model for risk stratification and treatment planning. *Acta Oto-Laryngologica*, 143(3), 206-214. <https://doi.org/10.1080/00016489.2023.2172208>
14. Adeoye, J., Tan, J. Y., Choi, S. W., & Thomson, P. (2021). Prediction models applying machine learning to oral cavity cancer outcomes: A systematic review. *International Journal of Medical Informatics*, 154, 104557.
15. Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J.P. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science & Technology*, 9(2), 14. <https://doi.org/10.1167/tvst.9.2.14>
16. Schwendicke, F., Samek, W., & Krois, J. (2020). Artificial intelligence in dentistry: Chances and challenges. *Journal of Dental Research*. <https://doi.org/10.1177/0022034520916632>
17. Chu CS, Lee NP, Adeoye J, Thomson P, Choi S-W. Machine learning and treatment outcome prediction for oral cancer. *J Oral Pathol Med*. 2020;00:1-9. <https://doi.org/10.1111/jop.13089>
18. Surveillance, Epidemiology, and End Results Program. Cancer Stat Facts: Oral Cavity and Pharynx Cancer. <https://seer.cancer.gov/statfacts/html/oralcav.html>
19. Spiotto MT, Jefferson GD, Wenig B, Markiewicz MR, Weichselbaum RR, Koshy M. Survival outcomes for postoperative chemoradiation in intermediate-risk oral tongue cancers. *Head & Neck*. 2017; 39: 2537-2548.

<https://doi-org.libproxy.temple.edu/10.1002/hed.24932>

20. Chiesa-Estomba, C. M., Grafia, M., Medela, A., Sistiaga-Suarez, J. A., Lechien, J. R., Calvo-Henriquez, C., Mayo-Yanez, M., Vaira, L.A., Grammatica, A., Cammaroto, G., Ayad, T., & Fagan, J. J. (2022). Machine learning algorithms as a computer-assisted decision tool for oral cancer prognosis and management decisions: A systematic review. *S. Karger AG, Basel*.
21. Du, M., Haag, D. G., Lynch, J. W., & Mittinty, M. N. (2020). Comparison of the tree-based machine learning algorithms to Cox regression in predicting the survival of oral and pharyngeal cancers: Analyses based on SEER database. *Cancers, 12* (10), 2802. <https://doi.org/10.3390/cancers12102802>
22. Kwak, M. S., Eun, Y.-G., Lee, J.-W., & Lee, Y. C. (2021). Development of a machine learning model for the prediction of nodal metastasis in early T classification oral squamous cell carcinoma: A SEER-based population study. *Head & Neck, 43*(8), 2316-2324. <https://doi.org/10.1002/hed.26700>
23. Sarode, G., Maniyar, N., & Sarode, S. C. (2020). Epidemiologic aspects of oral cancer. *Disease-a-Month*. <https://doi.org/10.1016/j.disamonth.2020.100988>
24. Hsu, P.-K., Huang, C.-S., Wang, B.-Y., Wu, Y.-C., & Hsu, W.-H. (2014). Survival benefits of postoperative chemoradiation for lymph node-positive esophageal squamous cell carcinoma. *The Annals of Thoracic Surgery, 97*(5), 1734-1741. <https://doi.org/10.1016/j.athoracsur.2013.12.041>
25. Cassidy, R. J., Switchenko, J. M., Jegadeesh, N., Sayan, M., Ferris, M. J., Eaton, B. R., Higgins, K. A., Wadsworth, J. T., Magliocca, K. R., Saba, N. F., & Beitler, J. J. (2017). Association of lymphovascular space invasion with locoregional failure and survival in patients with node-negative oral tongue cancers. *JAMA Otolaryngology-Head & Neck Surgery, 143*(4), 382-388. <https://doi.org/10.1001/jamaoto.2016.3795>
26. SEER\*Stat Variables Dictionary. (2020). *National Cancer Institute*. Retrieved April 2021, from <https://seer.cancer.gov/tools/variables/>
27. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8-

17. <https://doi.org/10.1016/j.csbj.2014.11.005>
28. Peng, J., Lu, Y., Chen, L., Qiu, K., Chen, F., Liu, J., Xu, W., Zhang, W., Zhao, Y., Yu, Z., & Ren, J. (2022). The prognostic value of machine learning techniques versus Cox regression model for head and neck cancer. *Methods*, 205, 123-132. <https://doi.org/10.1016/j.ymeth.2022.07.001>
29. Alabi, R. O., Makitie, A. A., Pirinen, M., Elmusrati, M., Leiva, I., & Almangush, A. (2021). Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer. *International Journal of Medical Informatics*, 145, 104313. <https://doi.org/10.1016/j.ijmedinf.2020.104313>
30. Du, M., Haag, D. G., Lynch, J. W., & Mittinty, M. N. (2020). Comparison of the tree-based machine learning algorithms to Cox regression in predicting the survival of oral and pharyngeal cancers: Analyses based on SEER database. *Cancers*, 12\*(10), 2802. <https://doi.org/10.3390/cancers12102802>
31. Tewari, P., Kashdan, E., Walsh, C., Martin, C. M., Parnell, A. C., & O'Leary, J. J. (2021). Estimating the conditional probability of developing human papilloma virus related oropharyngeal cancer by combining machine learning and inverse Bayesian modelling. *PLoS Computational Biology*, 17\*(8), e1009289. <https://doi.org/10.1371/journal.pcbi.1009289>
32. Kwak, M. S., Eun, Y.-G., Lee, J.-W., & Lee, Y. C. (2021). Development of a machine learning model for the prediction of nodal metastasis in early T classification oral squamous cell carcinoma: SEER-based population study. *Head & Neck*, 43 (8), 2316-2324. <https://doi.org/10.1002/hed.26700>
33. Fatapour, Y., Abiri, A., Kuan, E. C., & Brody, J.P. (2023). Development of a machine learning model to predict recurrence of oral tongue squamous cell carcinoma. *Cancers*, 15 (10), 2769. <https://doi.org/10.3390/cancers15102769>
34. Adeoye J, Tan N, Choi SW, Thomson P. Prediction models applying machine learning to oral cavity cancer outcomes: A systematic review. *Int J Med Inform.* 2021 Oct;154:104557. doi: 10.1016/j.ijmedinf.2021.104557. Epub 2021 Aug 18. PMID: 34455119.
35. Ahmed N, Abbasi MS, Zuberi F, Qamar W, Halim MSB, Maqsood A, Alam MK. Artificial Intelligence Techniques: Analysis, Application, and Outcome in Dentistry-A Systematic Review. *Biomed Res Int.* 2021 Jun 22;2021:9751564. doi:

10.1155/2021/9751564. PMID: 34258283; PMCID: PMC8245240.

36. Ellington TD, Henley SJ, Senkomago V, O'Neil ME, Wilson RJ, Singh S, Thomas CC, Wu M, Richardson LC. Trends in Incidence of Cancers of the Oral Cavity and Pharynx - United States 2007-2016. *MMWR Morb Mortal Wkly Rep.* 2020 Apr 17;69(15):433-438. doi: 10.1155/2021/9751564. PMID: 32298244; PMCID: PMC7755056.
37. Montero PH, Patel SG. Cancer of the oral cavity. *Surg Oncol Clin N Am.* 2015 Jul;24(3):491-508. doi: 10.1016/j.soc.2015.03.006. Epub 2015 Apr 15. PMID: 25979396; PMCID: PMC5018209.
38. Campo, E., Swerdlow, S. H., Harris, N. L., Pileri, S., Stein, H., & Jaffe, E. S. (2008). The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. *Blood*, *112*(19), 5019-5032. <https://doi.org/10.1182/blood-2011-02-320284>