ISSN 2218-4333 (online)

# World Journal of *Clinical Oncology*

*World J Clin Oncol* 2020 November 24; 11(11): 854-967

Published by Baishideng Publishing Group Inc

W J C O

World Journal of
Clinical Oncology

## Contents

Monthly Volume 11 Number 11 November 24, 2020

### FRONTIER

### REVIEW

### MINIREVIEWS

### ORIGINAL ARTICLE

#### Retrospective Study

#### Prospective Study

### CASE REPORT

## ABOUT COVER

Editorial board member of *World Journal of Clinical Oncology*, Dr. Fabrício Freire de Melo is a Professor at the Multidisciplinary Institute of Health of the Federal University of Bahia, Brazil. He undertook his postgraduate training at the Faculty of Medicine of the Federal University of Minas Gerais, where he received his Master's degree (2007) and PhD (2011), and completed his Postdoctoral Fellowship (2013) in Microbiology. His ongoing research interests involve the host-pathogen interactions in *Helicobacter pylori* gastric infection and the features associated with development of duodenal ulcer and gastric cancer. In addition, he has been investigating diagnostic methods for and the immune response in COVID-19. Currently, he serves as Research and Extension Coordinator of the Multidisciplinary Institute of Health of the Federal University of Bahia. (L-Editor: Filipodia)

## AIMS AND SCOPE

The primary aim of *World Journal of Clinical Oncology* (*WJCO*, *World J Clin Oncol*) is to provide scholars and readers from various fields of oncology with a platform to publish high-quality basic and clinical research articles and communicate their research findings online.

*WJCO* mainly publishes articles reporting research results and findings obtained in the field of oncology and covering a wide range of topics including art of oncology, biology of neoplasia, breast cancer, cancer prevention and control, cancer-related complications, diagnosis in oncology, gastrointestinal cancer, genetic testing for cancer, gynecologic cancer, head and neck cancer, hematologic malignancy, lung cancer, melanoma, molecular oncology, neurooncology, palliative and supportive care, pediatric oncology, surgical oncology, translational oncology, and urologic oncology.

## INDEXING/ABSTRACTING

The *WJCO* is now abstracted and indexed in PubMed, PubMed Central, Emerging Sources Citation Index (Web of Science), China National Knowledge Infrastructure (CNKI), China Science and Technology Journal Database (CSTJ), and Superstar Journals Database.

## RESPONSIBLE EDITORS FOR THIS ISSUE

**Production Editor:** *Li-Li Wang*; **Production Department Director:** *Yun-Xiaojian Wu*; **Editorial Office Director:** *Jia-Ping Yan*.

**NAME OF JOURNAL**
*World Journal of Clinical Oncology*

**ISSN**
ISSN 2218-4333 (online)

**LAUNCH DATE**
November 10, 2010

**FREQUENCY**
Monthly

**EDITORS-IN-CHIEF**
Hiten RH Patel

**EDITORIAL BOARD MEMBERS**
https://www.wjgnet.com/2218-4333/editorialboard.htm

**PUBLICATION DATE**
November 24, 2020

**COPYRIGHT**
© 2020 Baishideng Publishing Group Inc

**INSTRUCTIONS TO AUTHORS**
https://www.wjgnet.com/bpg/gerinfo/204

**GUIDELINES FOR ETHICS DOCUMENTS**
https://www.wjgnet.com/bpg/GerInfo/287

**GUIDELINES FOR NON-NATIVE SPEAKERS OF ENGLISH**
https://www.wjgnet.com/bpg/gerinfo/240

**PUBLICATION ETHICS**
https://www.wjgnet.com/bpg/GerInfo/288

**PUBLICATION MISCONDUCT**
https://www.wjgnet.com/bpg/gerinfo/208

**ARTICLE PROCESSING CHARGE**
https://www.wjgnet.com/bpg/gerinfo/242

**STEPS FOR SUBMITTING MANUSCRIPTS**
https://www.wjgnet.com/bpg/GerInfo/239

**ONLINE SUBMISSION**
https://www.f6publishing.com

W J C O World Journal of Clinical Oncology

Submit a Manuscript: https://www.f6publishing.com

World J Clin Oncol 2020 November 24; 11(11): 918-934

DOI: 10.5306/wjco.v11.i11.918

ISSN 2218-4333 (online)

ORIGINAL ARTICLE

Retrospective Study

# Artificial intelligence in dentistry: Harnessing big data to predict oral cancer survival

Man Hung, Jungweon Park, Eric S Hon, Jerry Bounsanga, Sara Moazzami, Bianca Ruiz-Negrón, Dawei Wang

**ORCID number:** Man Hung 0000-0003-2827-3740; Jungweon Park 0000-0001-7930-6026; Eric S Hon 0000-0002-8779-4397; Jerry Bounsanga 000-0001-6852-4650; Sara Moazzami 0000-0003-2403-3141; Bianca Ruiz-Negrón 0000-0001-6354-1582; Dawei Wang 0000-0003-3842-4258.

**Author contributions:** Hung M and Hon ES contributed to study conception; Hung M provided study supervision; Hung M and Wang D contributed to research design, data analysis, visualization and results interpretation; Hung M, Hon ES and Bounsanga J contributed to data acquisition; Hung M, Park J, Moazzami S, Ruiz-Negrón B and Wang D contributed to manuscript drafting; Hung M, Park J, Hon ES, Bounsanga J and Wang D contributed to manuscript revision; all authors approved the final version of the manuscript.

**Institutional review board statement:** This is not a human subject research study. Per the United States federal regulations (45 CFR 46, category 4), this study is deemed exempt and does not require review from Institutional Review Board since the data were deidentified and publicly available.

**Informed consent statement:** This

**Man Hung, Jungweon Park, Sara Moazzami,** College of Dental Medicine, Roseman University of Health Sciences, South Jordan, UT 84095, United States

**Man Hung,** Department of Orthopaedic Surgery Operations, University of Utah, Salt Lake City, UT 84108, United States

**Man Hung,** College of Social Work, University of Utah, Salt Lake City, UT 84112, United States

**Man Hung,** Division of Public Health, University of Utah, Salt Lake City, UT 84108, United States

**Man Hung,** Department of Educational Psychology, University of Utah, Salt Lake City, UT 84109, United States

**Eric S Hon,** Department of Economics, University of Chicago, Chicago, IL 60637, United States

**Jerry Bounsanga,** Research Section, Utah Medical Education Council, Salt Lake City, UT 84102, United States

**Bianca Ruiz-Negrón,** College of Social and Behavioral Sciences, University of Utah, Salt Lake City, UT 84112, United States

**Dawei Wang,** Data Analytics Unit, Walmart Inc., Bentonville, AR 72716, United States

**Corresponding author:** Man Hung, PhD, Professor, Research Dean, College of Dental Medicine, Roseman University of Health Sciences, 10894 S River Front Parkway, South Jordan, UT 84095, United States. mhung@roseman.edu

## Abstract

*BACKGROUND*

Oral cancer is the sixth most prevalent cancer worldwide. Public knowledge in oral cancer risk factors and survival is limited.

*AIM*

To come up with machine learning (ML) algorithms to predict the length of survival for individuals diagnosed with oral cancer, and to explore the most important factors that were responsible for shortening or lengthening oral cancer survival.

is not a human subject research study. This study used secondary data that were already collected and were publicly available online. Therefore, signed informed consent form is not relevant.

**Conflict-of-interest statement:** The authors declare that there is no conflict of interest regarding this work.

**Data sharing statement:** The data supporting the findings of this study can be accessed at: https://seer.cancer.gov/.

**Manuscript source:** Unsolicited manuscript

**Specialty type:** Oncology

**Country/Territory of origin:** United States

**Peer-review report's scientific quality classification**

Grade A (Excellent): A
Grade B (Very good): 0
Grade C (Good): 0
Grade D (Fair): 0
Grade E (Poor): 0

**Received:** June 24, 2020
**Peer-review started:** June 24, 2020
**First decision:** September 18, 2020
**Revised:** October 6, 2020
**Accepted:** October 20, 2020
**Article in press:** October 20, 2020
**Published online:** November 24, 2020

**P-Reviewer:** Santos-García G
**S-Editor:** Chen XF

***METHODS***

We used the Surveillance, Epidemiology, and End Results database from the years 1975 to 2016 that consisted of a total of 257880 cases and 94 variables. Four ML techniques in the area of artificial intelligence were applied for model training and validation. Model accuracy was evaluated using mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), $R^2$ and adjusted $R^2$.

***RESULTS***

The most important factors predictive of oral cancer survival time were age at diagnosis, primary cancer site, tumor size and year of diagnosis. Year of diagnosis referred to the year when the tumor was first diagnosed, implying that individuals with tumors that were diagnosed in the modern era tend to have longer survival than those diagnosed in the past. The extreme gradient boosting ML algorithms showed the best performance, with the MAE equaled to 13.55, MSE 486.55 and RMSE 22.06.

***CONCLUSION***

Using artificial intelligence, we developed a tool that can be used for oral cancer survival prediction and for medical-decision making. The finding relating to the year of diagnosis represented an important new discovery in the literature. The results of this study have implications for cancer prevention and education for the public.

**Key Words:** Oral cancer survival; Machine learning; Artificial intelligence; Dental medicine; Public health; Surveillance, Epidemiology, and End Results; Quality of life

**Core Tip:** Oral cancer is the sixth most prevalent cancer worldwide. The goal of this study was to come up with machine learning algorithms to predict the length of oral cancer survival and to explore the most important factors that were responsible for it. Age at diagnosis, primary cancer site, tumor size and year of diagnosis were found to be the most important factors predictive of oral cancer survival. Year of diagnosis represents an important new discovery in the literature. Using artificial intelligence, we developed a tool that can be used for oral cancer survival prediction and for medical decision making.

**Citation:** Hung M, Park J, Hon ES, Bounsanga J, Moazzami S, Ruiz-Negrón B, Wang D. Artificial intelligence in dentistry: Harnessing big data to predict oral cancer survival. *World J Clin Oncol* 2020; 11(11): 918-934
**URL:** https://www.wjgnet.com/2218-4333/full/v11/i11/918.htm
**DOI:** https://dx.doi.org/10.5306/wjco.v11.i11.918

## INTRODUCTION

To minimize the occurrence of oral cancer and improve one's quality of life, it is imperative to conduct screenings for early detection of head and neck carcinomas (HNC) on all high-risk dental patients. HNC, which is the umbrella term that includes oral cancer, are often located within the oral and nasal cavities, upper/lower pharynx, larynx, and the maxillary sinus[1-3]. Early screenings for identification of dysplastic tissue in the head and neck region are within the scope of care of the dental health providers. Oral cancer may be curable if detected early[4]. However, more than one-half of all oral and pharyngeal cancers in the United States were detected at late stages[4,5], thus the overall United States five-year survival rate for oral cancer was only 52 percent[6]. In 2012, there were 145000 deaths in the United States attributed to oral cancers[2]. Throughout the world, approximately 563826 diagnoses of oral cancer were reported, rendering it the sixth most common type of cancer in the world[7-10]. Although there is a downward trend in oral cancer incidence due to the rising awareness in the risks associated with tobacco use and alcohol consumption in the United States, a

**L-Editor:** A
**P-Editor:** Wang LL

general lack of public awareness of the symptoms and other risk factors of oral cancer remains[4]. In 2016, a total of 48330 oral cavity and oropharyngeal cancer incidents were reported[11], and an increase of 225% of human papillomavirus-related oropharynx cancer was recorded[11]. Altogether, the need to address additional risk factors and increases in early screenings of oral cancers are key factors to improving cancer survival[12].

Among those with an oral cancer diagnosis, stage of tumor at time of diagnosis and treatment have been associated with survival[10,13]. Specifically, in a study by Sargeran *et al*[13] the survival rates were higher in patients with stages I or II cancer than those with stage III cancer at the time of the diagnosis. They further concluded that patients who had undergone radiotherapy alone had a lower survival rate than patients with a combination of surgery and radiotherapy, and that age and sex were not associated with survival. However, Warnakulasuriya *et al*[10] found that younger age was associated with higher 5-year relative survival rate.

Additionally, race has been associated with varying level of survival rates. A study using 1973-2002 data from Surveillance, Epidemiology, and End Results (SEER-18)[14] by Shiboski *et al*[15] revealed that the stage at diagnosis was related to 5-year relative survival rate among Whites and Blacks. The results indicated that Blacks had a significantly higher rate of cancer, mainly located on the tongue, with tumors larger than 4 cm in diameter at the time of diagnosis. Black men experienced lower 5-year relative survival rates compared to White men, especially for tongue cancer. Shiboski *et al*[15] explained that the differences in survivals across different races may be due to differences in access to, and utilization of healthcare services.

Due to the limited understanding of the disparities seen across cancer survivors and public knowledge on risk factors and symptoms, investigators in the past have suggested for primary care providers to put greater weight on initial screening and comprehensive soft-tissue exams[15]. Having a tool to accurately predict the survival time of oral cancer patients could help regulate the effects of psychological distress on physical and mental health outcomes after diagnosis. Medical decision-making tools based on fuzzy and soft set theories and artificial intelligence are effective for determination of cancer survival and enhancing disease awareness[16]. Awareness of the disease can lessen the burden of the disease on the survivors and their caretakers, and assist with medical and dental decision-making moving forward. The main purpose of this study was to apply artificial intelligence to build a model to predict the length of survival for those diagnosed with oral cancer as accurately and precisely as possible based on 40 plus years-worth of data representative of the United States' population. The secondary purpose was to explore the most important factors that were influencing the longevity of oral cancer survival.

## MATERIALS AND METHODS

### Data

Data from the SEER-18 database[14] were used to conduct this study. The SEER-18 database is a population-based registry that contains cancer-related data on individuals diagnosed with cancer from hospitals and laboratories in the United States[14]. The SEER-18 database does not contain data from Louisiana during hurricanes Katrina and Rita from July to December in 2005[14]. Institutional review board approval was not required for this study since the SEER-18 data were deidentified and publicly available online. The data that support the findings of this study are openly available at https://seer.cancer.gov/.

Oral cancer cases from the years 1975 to 2016 in the SEER-18 database[17] were identified by the International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) site codes (https://training.seer.cancer.gov/head-neck/abstract-code-stage/codes.html)[18]. Table 1 contains a list of all ICD-O-3 site codes that were identified for the oral cancer cases utilized in this study[18-20].

### Analytical approach

The outcome of interest for this study was oral cancer survival time. Survival time represented the time of survival in months from the date of cancer diagnosis to the date of last contact[21,22].

Descriptive statistics of demographics and cancer characteristics (such as primary site, tumor size, laterality, *etc.*) were analyzed. Prediction of oral cancer survival time was modeled by using four machine learning (ML) algorithms: linear regression, decision tree, random forest, and extreme gradient boosting (XGBoost). ML is a

**Table 1 Number of oral cancer cases from various anatomical sites**

| ICD-O-3 codes | Sites | Number of cases |
|---|---|---|
| C000 | External upper lip | 413 |
| C001 | External lower lip | 2444 |
| C002 | External lip, NOS | 92 |
| C003 | Mucosa of upper lip | 104 |
| C004 | Mucosa of lower lip | 567 |
| C005 | Mucosa of lip, NOS | 29 |
| C006 | Commissure of lip | 85 |
| C008 | Overlapping lesion of lip | 46 |
| C009 | Lip, NOS (excludes skin of lip C44.0) | 153 |
| C019 | Base of tongue, NOS | 10840 |
| C020 | Dorsal surface of tongue, NOS | 652 |
| C021 | Border of tongue | 2632 |
| C022 | Ventral surface of tongue, NOS | 1688 |
| C023 | Anterior 2/3 of tongue, NOS | 2807 |
| C024 | Lingual tonsil | 170 |
| C028 | Overlapping lesion of tongue | 581 |
| C029 | Tongue, NOS | 3050 |
| C030 | Upper gum | 821 |
| C031 | Lower gum | 1680 |
| C039 | Gum, NOS | 210 |
| C040 | Anterior floor of mouth | 1362 |
| C041 | Lateral floor of mouth | 352 |
| C048 | Overlapping lesion of floor of mouth | 136 |
| C049 | Floor of mouth, NOS | 2284 |
| C050 | Hard palate | 1155 |
| C051 | Soft palate, NOS (excludes nasopharyngeal surface of soft palate C11.3) | 1301 |
| C052 | Uvula | 180 |
| C058 | Overlapping lesion of palate | 206 |
| C059 | Palate, NOS | 154 |
| C060 | Cheek mucosa | 1787 |
| C061 | Vestibule of mouth | 134 |
| C062 | Retromolar area | 1413 |
| C068 | Overlapping lesion of other and unspecified parts of mouth | 142 |
| C069 | Mouth, NOS | 487 |
| C079 | Parotid gland | 7111 |
| C080 | Submandibular gland | 1149 |
| C081 | Sublingual gland | 94 |
| C088 | Overlapping lesion of major salivary glands | 6 |
| C089 | Major salivary gland, NOS (excludes minor salivary gland, NOS C06.9) | 287 |
| C090 | Tonsillar fossa | 1735 |
| C091 | Tonsillar pillar | 888 |

| | | |
|---|---|---|
| C098 | Overlapping lesion of tonsil | 109 |
| C099 | Tonsil, NOS (excludes lingual tonsil C02.4 and pharyngeal tonsil C11.1) | 9521 |
| C100 | Vallecula | 282 |
| C101 | Anterior surface of epiglottis | 88 |
| C102 | Lateral wall of oropharynx | 184 |
| C103 | Posterior wall of oropharynx | 246 |
| C104 | Branchial cleft (site of neoplasm) | 37 |
| C108 | Overlapping lesion of oropharynx | 277 |
| C109 | Oropharynx, NOS | 940 |
| C129 | Pyriform sinus | 1707 |
| C130 | Postcricoid region | 78 |
| C131 | Hypopharyngeal aspect of aryepiglottic fold, NOS (excludes laryngeal aspect of aryepiglottic fold C32.1) | 214 |
| C132 | Posterior wall of hypopharynx | 250 |
| C138 | Overlapping lesion of hypopharynx | 113 |
| C139 | Hypopharynx, NOS | 816 |
| C739 | Thyroid gland | 111425 |

ICD-O-3 codes: International Classification of Diseases for Oncology, 3rd ed; NOS: Not otherwise specified.

computer algorithm-based method that can efficiently detect relationships between variables with unrecognizable trends in large and complex data. The process takes into account historical trends to come up with models in predicting outcome of interest ( *e.g.*, oral cancer survival time), and then validates the models with actual or current data. The performance of the various models from the validation process will be compared, and the more parsimonious model with better performance is generally the preferred model. The ML techniques included in this study were chosen due to their ability to prevent over-fitting, being commonly used in similar studies, and their ease of interpretation in medical settings. To compare the different techniques, model accuracy was evaluated using mean absolute error (MAE)[1,13], mean squared error (MSE), root mean squared error (RMSE), $R^2$ and adjusted $R^2$. All analyses were conducted using Python 3.7.4 (Python Software Foundation).

There was a total of 257880 oral cancer cases and 94 variables (*i.e.*, features) in the dataset. Cases with missing data on the outcome variable (*i.e.*, oral cancer survival time) were dropped, and responses that were marked as not applicable were excluded. All variables with more than 40% of missing values were also excluded. Further data processing was conducted to remove null features, constant features (*i.e.*, features with same values for the outcome), quasi-constant features (*i.e.*, features with variance less than 0.01), and highly correlated features (*i.e.*, features with correlation higher than 0.9). These features were removed prior to data analysis as they would not contribute to the prediction of outcome and can often cause errors in the prediction. Outliers were detected by plotting distributions of each variable and they were replaced by mean, mode, and quantile as appropriate. Features with more than 90% the values that were the same were dropped.

To avoid the impracticality of including too many variables, further feature selection was performed using random forest. We aimed to narrow down the variables as much as possible without losing prediction accuracy. The random forest model showed that many features are of little importance (Figure 1). We dropped 7 features that were of less importance in terms of their importance scores, and a step backward feature selection method with random forest was then applied to select the best number of features. The cross-validation scores were then plotted (Figure 2) and the most important 10 features were kept to create a parsimony model. The cross-validation scores did not change much even after deleting the less significant features. The selected 10 features were: Year of diagnosis; primary site; age at diagnosis; CS tumor size; CS extension; CS lymph nodes eval; RX Summ-surg prim site; derived AJCC stage group; site recode ICD-O-3/WHO 2008; and month of diagnosis.

The final dataset used for model prediction from linear regression, decision tree,

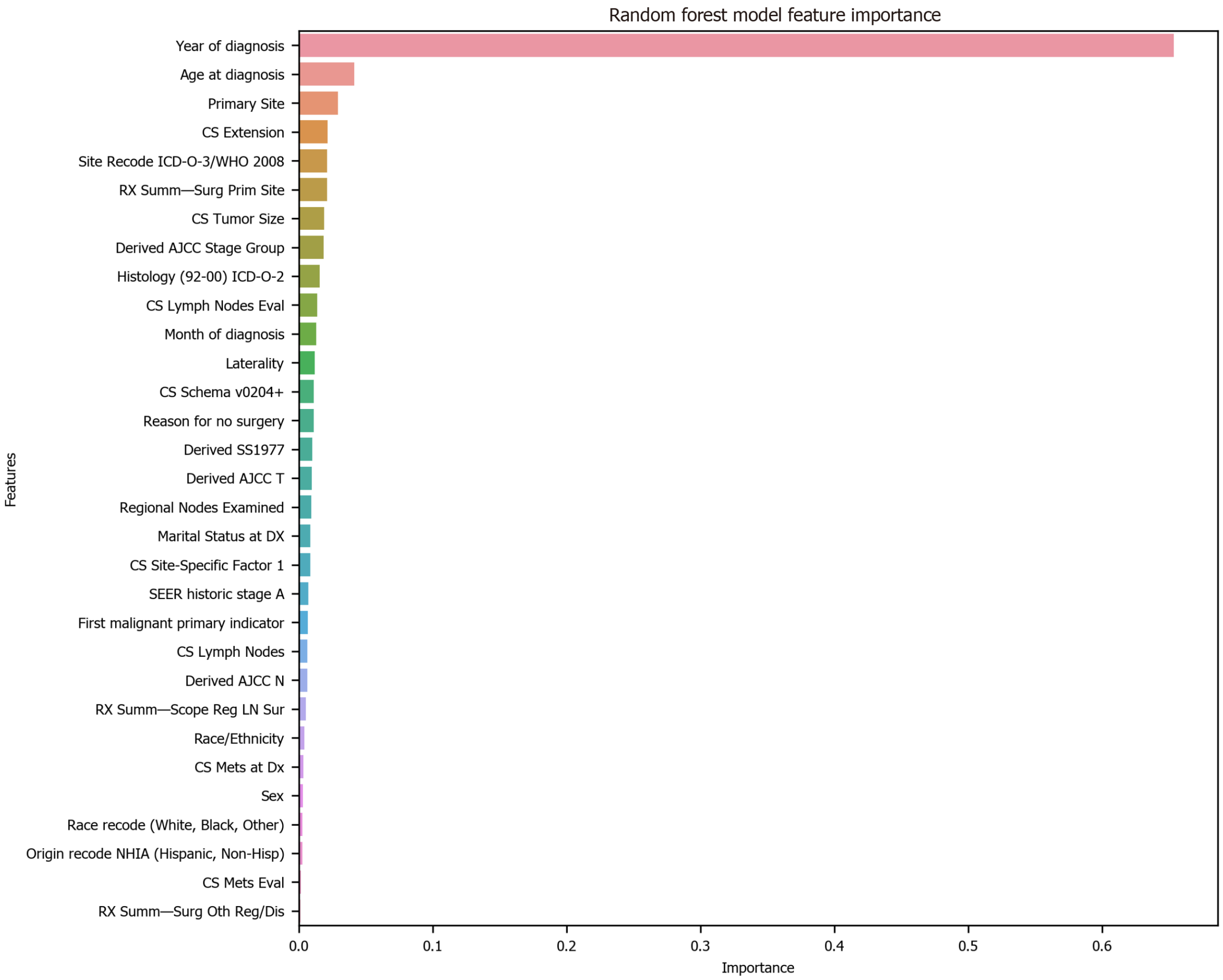


**Figure 1 Feature selection using random forest.** CS: Coding system; ICD-O-3: International Classification of Diseases for Oncology, 3rd ed; WHO: World Health Organization; AJCC: American Joint Committee on Cancer; SEER: Surveillance, Epidemiology, and End Results; LN: Lymph node.

random forest, and XGBoost had an effective sample size of 177714 cases with a total of 10 variables. Most of the values were categorized and given numerical code values. Table 2 lists all of the variables. Data were randomly split into training set and testing set. The training set contained 75% of the data and were used to build models. The testing set contained 25% of the data and were used to validate the models built from the training data. Detailed model parameter tuning set up is available upon request from the authors.

## RESULTS

There was a total of 177714 oral cancer cases included in the study, of which 63111 were oropharyngeal cancer cases and 114603 were laryngeal cancer cases. The nasopharyngeal cancer cases did not make it to the final sample since there was very few of these cases and all of them had a large number of missing values. Oropharynx cancer included anatomical positions at the base of tongue, lingual tonsil, soft palate, uvula, tonsil, orpharynx, Waldeyer ring, and histology sites[23]. Laryngeal cancer included areas at the larynx, which comprises of the epiglottis, supraglottis, vocal cord, glottis, and subglottis[24]. The sample consisted of 40.62% ($n$ = 72179) males. The average age at diagnosis was 54.6 years old (range: 0-109) (Figure 3). Nearly 40% of the sample were 60 years or older at the time of oral cancer diagnosis (Table 3).

Among the 10 features, several of them showed strong linear relation with survival time (Figure 4). Hence a linear regression model was used to predict outcome. The feature importance can be visualized in Figure 4 showing year of diagnosis as the most important variable. The performance of linear regression was MSE = 647.49, RMSE = 25.45, MAE = 18.21, $R^2$ = 0.620 and adjusted $R^2$ = 0.620 (Table 4).

**Table 2 List of all 10 variables included in the final machine learning model building and validation**

| Variables | Variable description |
|---|---|
| Age at diagnosis | This data item represents the age of the patient at diagnosis for this cancer. The code is three digits and represents the patient’s actual age in years |
| Year of diagnosis | The year of diagnosis is the year the tumor was first diagnosed by a recognized medical practitioner, whether clinically or microscopically confirmed |
| Month of diagnosis | The month of diagnosis is the month the tumor was first diagnosed by a recognized medical practitioner, whether clinically or microscopically confirmed |
| Primary site | This data item identifies the site in which the primary tumor originated. See the International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3)[18] for topography codes. The decimal point is eliminated |
| CS tumor size | Information on tumor size. Available for 2004-2015 diagnosis years. Earlier cases may be converted and new codes added which weren't available for use prior to the current version of CS. For more information, see http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage[19] |
| CS extension | Information on extension of the tumor. Available for 2004-2015 diagnosis years. Earlier cases may be converted and new codes added which weren't available for use prior to the current version of CS. For more information, see http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage[19] |
| CS lymph nodes eval | Available for 2004-2015, but not required for the entire timeframe. Will be blank in cases not collected. For more information, see http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage[19] |
| Derived AJCC stage group | This is the AJCC “Stage Group” component that is derived from CS detailed site-specific codes, using the CS algorithm, effective with 2004-2015 diagnosis years. See the CS site-specific schema for details (http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage)[19] |
| RX Summ-surg prim site | Surgery of primary site describes a surgical procedure that removes and/or destroys tissue of the primary site performed as part of the initial work-up or first course of therapy |
| Site recode ICD-O-3/WHO 2008 | A recode based on primary site and ICD-O-3 Histology in order to make analyses of site/histology groups easier. For example, the lymphomas are excluded from stomach and Kaposi and mesothelioma are separate categories based on histology. For more information, see http://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html[20] |

CS: Coding System; AJCC: American Joint Committee on Cancer; ICD-O-3: International Classification of Diseases for Oncology, 3rd ed; WHO: World Health Organization.

Decision tree regression, a ML method, was used to determine the top features (*i.e.*, variables) that were predictive of oral cancer survival time. Relative variable importance scores were computed to identify the top predictors. The usage of the decision tree regression was ideal as it doesn’t require linear relationship between features and target variable. Year of diagnosis was found as the most important variable (Figure 5). The performance of the decision tree was MSE = 538.30, RMSE = 23.20, MAE[1] = 14.45, $R^2$ = 0.681 and adjusted $R^2$ = 0.681 (Table 4).

Among the 10 features, several of them showed strong linear relation with survival time (Figure 4). Hence a linear regression model was used to predict outcome. The feature importance can be visualized in Figure 4 showing year of diagnosis as the most important variable. The performance of linear regression was MSE = 647.49, RMSE = 25.45, MAE = 18.21, $R^2$ = 0.620 and adjusted $R^2$ = 0.620 (Table 4).

Random forest method was also conducted to develop predictive model. It was appropriate for data with one strong predictor and some moderate predictors. The feature importance for random forest is shown in Figure 4 with year of diagnosis as the most important variable. The performance of the random forest was MSE = 489.58, RMSE = 22.13, MAE = 13.63, $R^2$ = 0.709 and adjusted $R^2$ = 0.709 (Table 4).

Finally, the XGBoost model was used. The performance of the XGBoost was MSE = 486.55, RMSE = 22.06, MAE = 13.55, $R^2$ = 0.711 and adjusted $R^2$ = 0.711 (Table 4). The feature importance for the XBoost model is presented in Figure 4 showing primary cancer site and year of diagnosis as the top two most important variables for prediction of oral cancer survival. Figure 6 presents a comparison of the prediction of oral cancer survival time from all models against the actual survival time. All model predictions were very similar and close to the actual outcomes. When the survival time was between 40 mo and 60 mo, the predictions were on target with the actual survival time. When it was under 40 mo, the predicted survival time for all models were slightly higher than the actual survival time. However, when it was over 60 mo, the predicted survival time for all models were slightly lower than the actual survival.

Table 3 Demographic characteristics of the sample (*n* = 177714)

| Variable | Mean | SD | Median | *n* | % |
|---|---|---|---|---|---|
| **Survival months/mo** | 60.35 | 40.98 | 54.00 | | |
| **Age at diagnosis/yr** | 54.62 | 16.10 | 55.00 | | |
| **Tumor size/(ID, cm)** | 22.56 | 21.74 | 19.00 | | |
| **Marital status** | | | | | |
| Single | | | | 35688 | 20.08 |
| Married | | | | 110480 | 62.17 |
| Separated | | | | 1746 | 0.98 |
| Divorced | | | | 16401 | 9.23 |
| Widowed | | | | 13055 | 7.35 |
| Unmarried or domestic partner | | | | 344 | 0.19 |
| **Sex** | | | | | |
| Male | | | | 72179 | 40.62 |
| Female | | | | 105535 | 59.38 |
| **Race** | | | | | |
| White | | | | 148556 | 83.60 |
| Black | | | | 16051 | 9.03 |
| Other | | | | 13107 | 7.38 |

SD: Standard deviation; ID: Diameter.

Table 4 Machine learning model performance

| Performance indicators | Linear regression | Decision tree | Random forest | XGBoost |
|---|---|---|---|---|
| MSE | 647.49 | 538.30 | 489.58 | 486.55 |
| RMSE | 25.45 | 23.20 | 22.13 | 22.06 |
| MAE | 18.21 | 14.45 | 13.63 | 13.55 |
| $R^2$ score | 0.620 | 0.681 | 0.709 | 0.711 |
| Adjusted $R^2$ score | 0.620 | 0.681 | 0.709 | 0.711 |

XGBoost: Extreme gradient boosting; MSE: Mean squared error; RMSE: Root mean squared error; MAE: Mean absolute error.

## DISCUSSION

The goal of this study was two-fold: (1) To build a ML model predictive of the length of survival for those diagnosed with oral cancer, and (2) To establish the most important factors that predict oral cancer survival. Our results showed that XGBoost was the best model in terms of accuracy. XGBoost's performance exceeded all other ML methods, with linear regression's performance slightly trailing behind all models. The average length of survival for all patients was 60.35 mo. Furthermore, age at diagnosis, primary cancer site, tumor size and year of diagnosis were the most important factors related to oral cancer survival. Year of diagnosis was consistently ranking as the top feature across all models. Year of diagnosis was not the number of years nor the amount of time since the tumor was initially diagnosed. Rather, year of diagnosis referred to the year when the tumor was first diagnosed, implying that individuals with tumors that were diagnosed in the modern era tend to have longer survival than those diagnosed in the past.

To our knowledge, this study is the first of its kind to use ML techniques to predict length of survival for those diagnosed with oral cancer. Previous research is consistent with some of our findings. Tumor size, specifically thickness among other tumor size

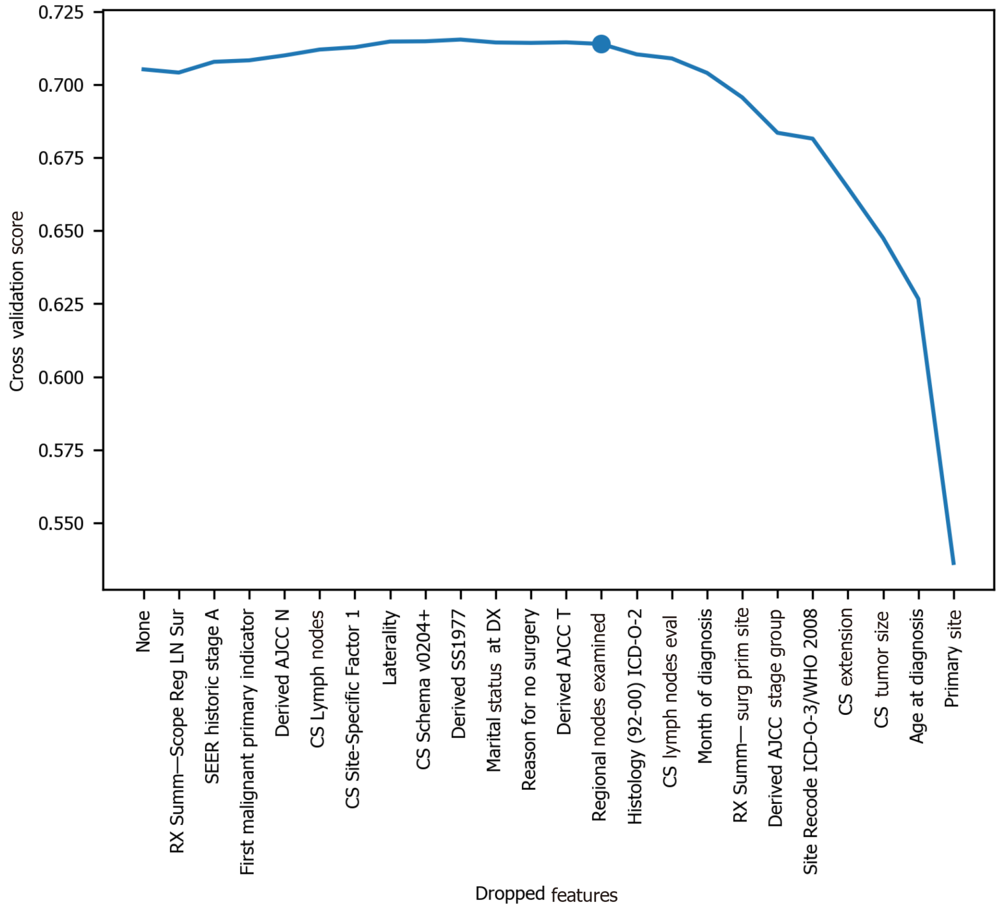


**Figure 2 Cross-validation score change for selecting optimal number of features.** LN: Lymph node; SEER: Surveillance, Epidemiology, and End Results; AJCC: American Joint Committee on Cancer; CS: Coding system; ICD-O-3: International Classification of Diseases for Oncology, 3rd ed; WHO: World Health Organization.

parameters, has been found to be a significant predictor of oral tongue carcinoma survival[25,26]. Younger patients with oral cavity squamous cell cancer[27] and squamous cell carcinoma of the oral tongue[28] have been found to have a higher survival rate in the past which is also consistent with our findings. For cases of squamous cell carcinoma of the oral tongue, a ten-year increase in age was associated with an 18% increase in risk of death[27,28]. However, year of diagnosis was a unique and novel predictive factor that has not been reported in the literature. Considering that our study included 40 plus years-worth of data and incorporated ML for precise prediction, this perhaps makes it possible for discovering new knowledge. It is possible that more recent year of diagnosis leads to the better survival outcomes due to improved oral cancer treatments and public awareness.

This study also revealed some conflicting findings that need further exploration. Although race and ethnicity have been identified as predictors to oral cancer survival in past literature[15], our study using recent data showed low importance of these features, so race and ethnicity were eventually dropped from the model. Given that our study included 40 plus year-worth of data and consisted of recent data, we may see that race and ethnicity are not associated with oral cancer survival over time. Improvements in access to, and utilization of healthcare services among race could also be reasons leading to no or low racial disparities in oral care in the 21st century. Additional large-scale studies using recent data are needed to evaluate these findings.

A primary limitation of this study was that the data did not include psychological factors that could explain survivors' quality of life. In a future study, we can explore other databases and incorporate surveys to explain the psychological state of oral cancer survivors and overall perspective on the disease. Over 50% of diagnosed oral cancer cases still remain a lethal disease annually[10], early detection and accessibility to regular head and neck examination is key.

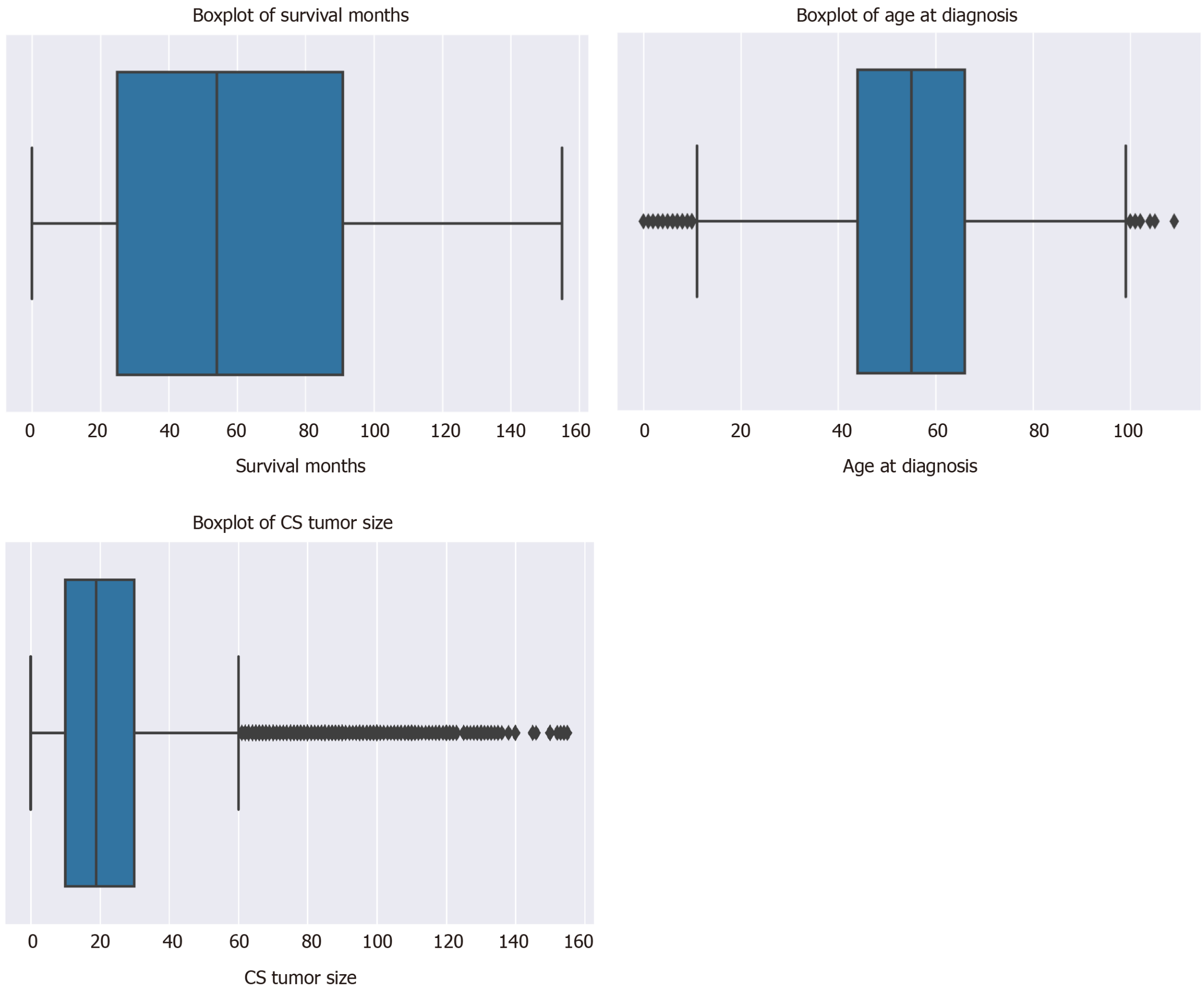
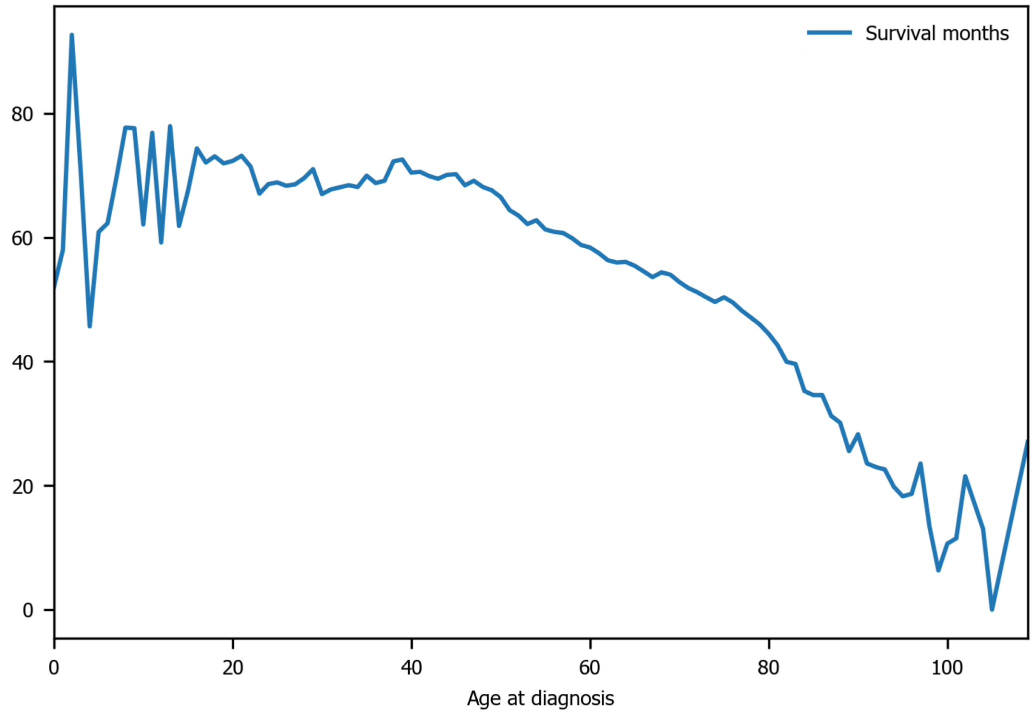
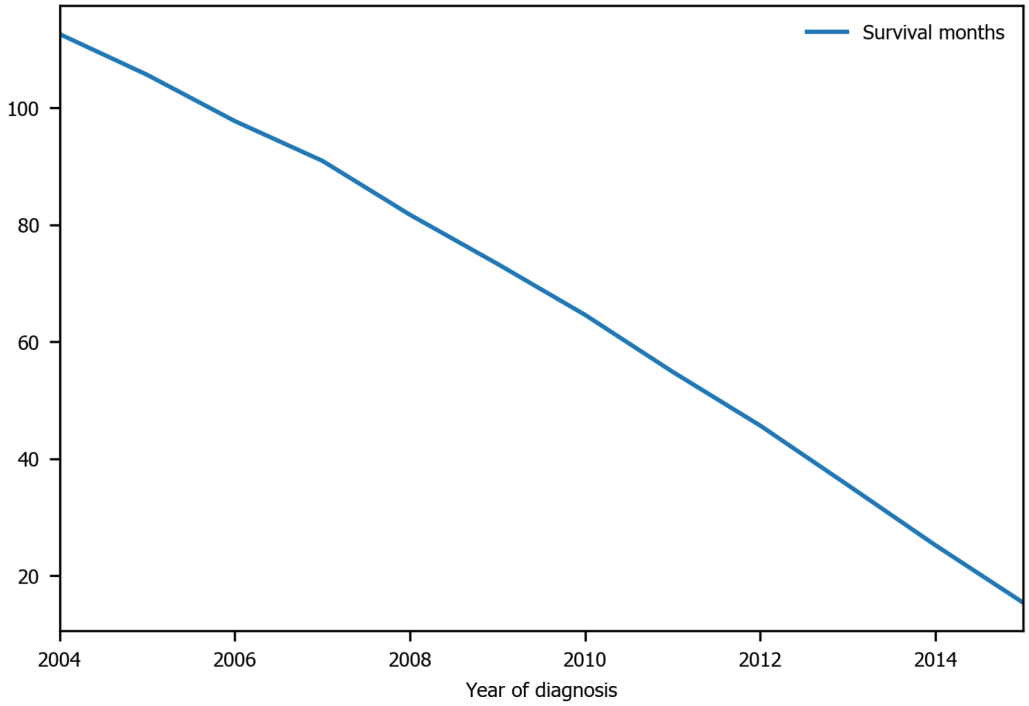


**Figure 3 Boxplots of sample characteristics.** CS: Coding system.

## CONCLUSION

This study is particularly important and appropriate for the field of dentistry as the prediction of oral cancer survival can assist dentists, patients and caregivers in disease management and treatment plan development. Identifying oral cancer and gaining a more in depth understanding of the length of survival for those diagnosed with oral cancer and establishing important factors that predict oral cancer survival will better equip health care providers on how to best manage such diagnoses. This study serves as a steppingstone for future exploration using ML and artificial intelligence to uncover the full potential for the management of oral cancers and to reduce healthcare disparities around the globe.


Survival months
80
60
40
20
0
0
20
40
60
80
100
Age at diagnosis


Survival months
100
80
60
40
20
2004
2006
2008
2010
2012
2014
Year of diagnosis

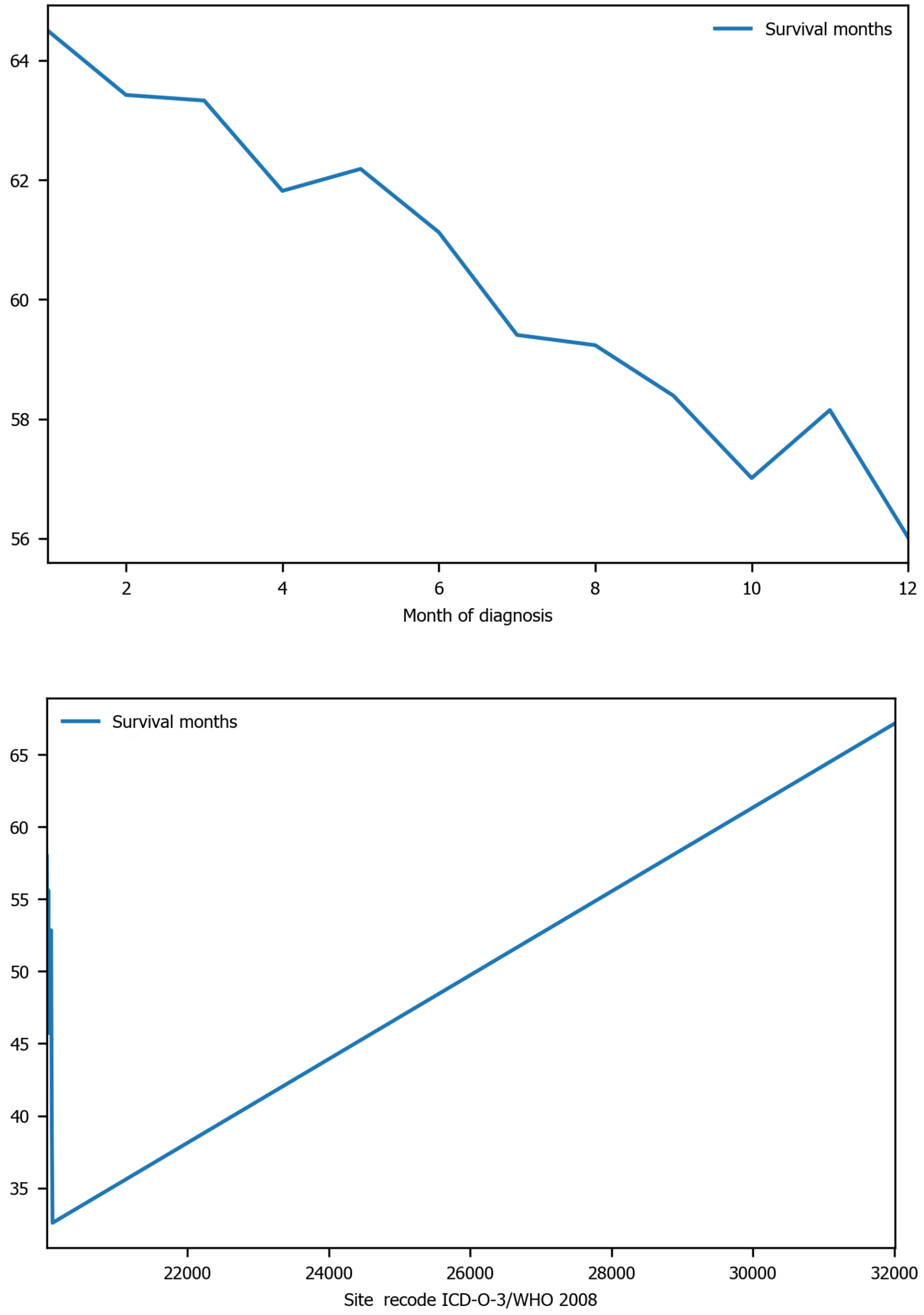
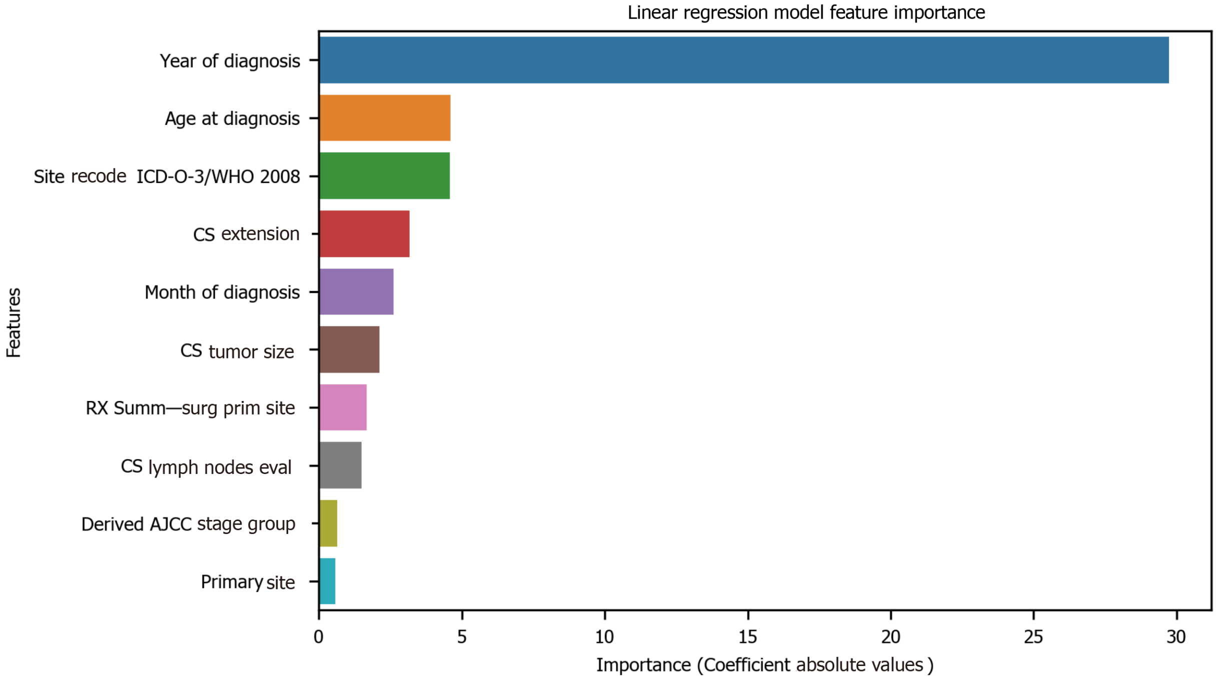
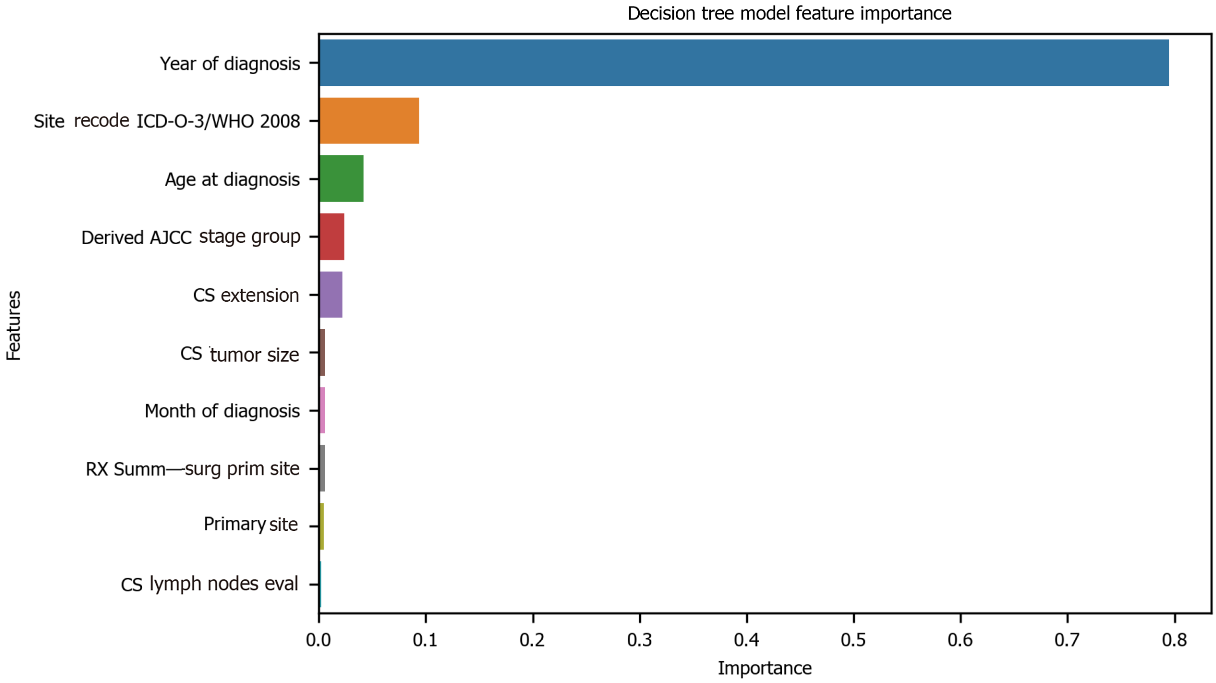


**Figure 4 Survival months shows strong linear relation with several variables: Age of diagnosis, year of diagnosis, month of diagnosis, and site recode ICD-O-3/WHO 2008.** ICD-O-3: International Classification of Diseases for Oncology, 3rd ed; WHO: World Health Organization.


Linear regression model feature importance
Features
Year of diagnosis
Age at diagnosis
Site recode ICD-O-3/WHO 2008
CS extension
Month of diagnosis
CS tumor size
RX Summ—surg prim site
CS lymph nodes eval
Derived AJCC stage group
Primary site
0
5
10
15
20
25
30
Importance (Coefficient absolute values )


Decision tree model feature importance
Features
Year of diagnosis
Site recode ICD-O-3/WHO 2008
Age at diagnosis
Derived AJCC stage group
CS extension
CS tumor size
Month of diagnosis
RX Summ—surg prim site
Primary site
CS lymph nodes eval
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Importance

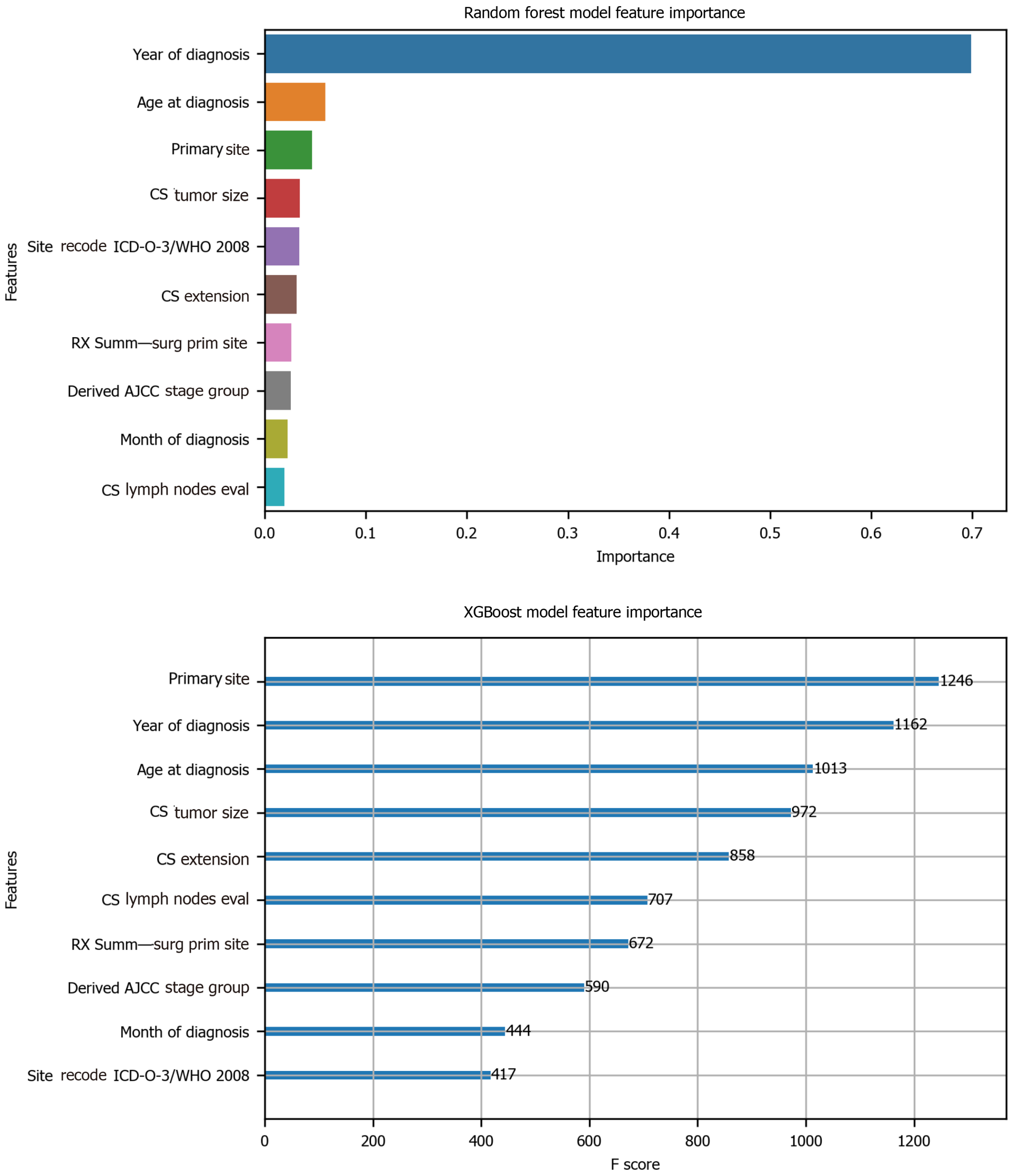


**Figure 5 Machine learning model feature importance.** ICD-O-3: International Classification of Diseases for Oncology, 3rd ed; WHO: World Health Organization; CS: Coding system; AJCC: American Joint Committee on Cancer.

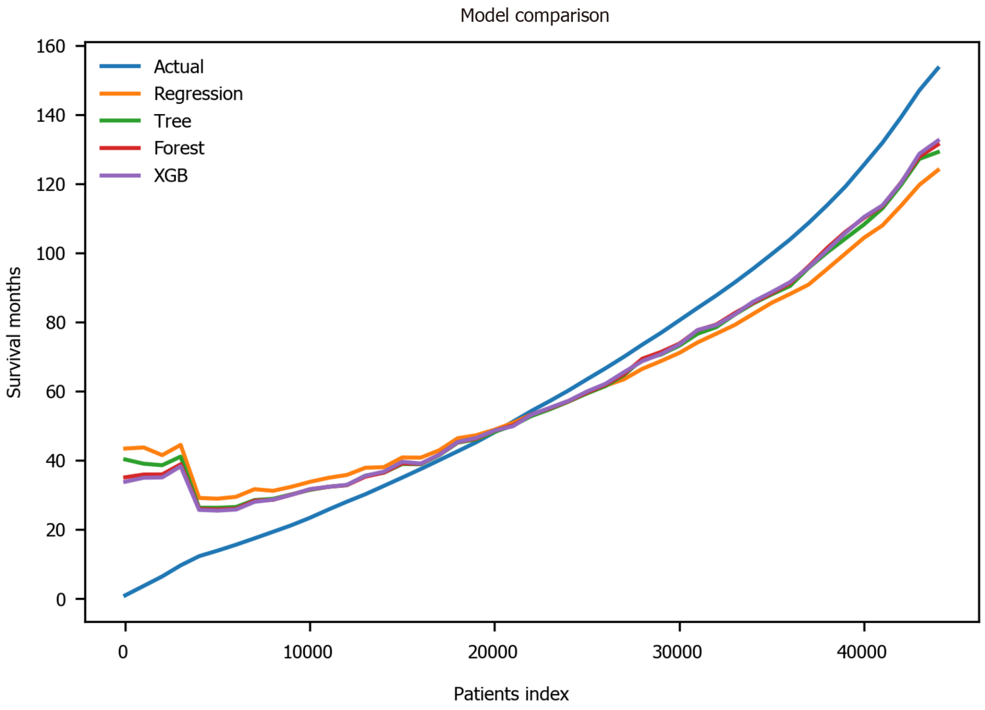


**Figure 6 Prediction comparison among different models.** Patient index refers to the rank after sorting by survival months. Actual: The actual survival outcome; XGB: Extreme gradient boosting.

## ARTICLE HIGHLIGHTS

### *Research background*

Oral cancer is highly prevalent in the world, yet there is a limited understanding of oral cancer risk factors and survival.

### *Research motivation*

To increase one's quality of life, it is important to be able to predict oral cancer survival.

### *Research objectives*

The objectives of this study were to build an accurate model to precisely predict the length of oral cancer survival and to explore the most important factors that determine the longevity of oral cancer survivors.

### *Research methods*

Oral cancer data were obtained from the years 1975 to 2016 in the Surveillance, Epidemiology, and End Results database. Methods from the field of artificial intelligence were applied to build and validate prediction models from 40+ years of oral cancer data representative of the United States' population.

### *Research results*

Age at diagnosis, primary cancer site, tumor size and year of diagnosis were the most important factors related to oral cancer survival. Individuals with tumors that were diagnosed in the modern era tend to have longer survival than those diagnosed in the past, which was a novel finding that had not been reported in the literature.

### *Research conclusions*

Machine learning algorithms were developed this study to predict the length of oral cancer survival that can be readily deployed to clinical settings.

### *Research perspectives*

This study was the first of its kind to use methods from artificial intelligence to examine the length of survival for individuals diagnosed with oral cancer. The outcome of this study has the potential to reduce healthcare disparities and improve the quality of life for oral cancer survivors and their friends and families around the world.

## ACKNOWLEDGEMENTS

The authors sincerely thank the Clinical Outcomes Research and Education at College of Dental Medicine, Roseman University of Health Sciences for supporting this study.

## REFERENCES

1 **Xu Z**, Neoh KG, Amaechi B, Kishen A. Monitoring bacterial-demineralization of human dentine by electrochemical impedance spectroscopy. *J Dent* 2010; **38**: 138-148 [PMID: 19804810 DOI: 10.1016/j.jdent.2009.09.013]

2 **Mourad M**, Jetmore T, Jategaonkar AA, Moubayed S, Moshier E, Urken ML. Epidemiological Trends of Head and Neck Cancer in the United States: A SEER Population Study. *J Oral Maxillofac Surg* 2017; **75**: 2562-2572 [PMID: 28618252 DOI: 10.1016/j.joms.2017.05.008]

3 **National Cancer Institute: Surveillance, Epidemiology, and End Results Program**. SEER*Stat Databases: November 2019 Submission; Incidence - SEER Research Data, 9 Registries, Nov 2019 Sub (1975-2017). 2019. Available from: https://seer.cancer.gov/data-software/documentation/seerstat/nov2019/

4 **Cruz GD**, Le Geros RZ, Ostroff JS, Hay JL, Kenigsberg H, Franklin DM. Oral cancer knowledge, risk factors and characteristics of subjects in a large oral cancer screening program. *J Am Dent Assoc* 2002; **133**: 1064-1071 [PMID: 12198985 DOI: 10.14219/jada.archive.2002.0330]

5 **McGurk M**, Chan C, Jones J, O'regan E, Sherriff M. Delay in diagnosis and its effect on outcome in head and neck cancer. *Br J Oral Maxillofac Surg* 2005; **43**: 281-284 [PMID: 15993279 DOI: 10.1016/j.bjoms.2004.01.016]

6 **Swango PA**. Cancers of the oral cavity and pharynx in the United States: an epidemiologic overview. *J Public Health Dent* 1996; **56**: 309-318 [PMID: 9089526 DOI: 10.1111/j.1752-7325.1996.tb02458.x]

7 **Ferlay J**, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015; **136**: E359-E386 [PMID: 25220842 DOI: 10.1002/ijc.29210]

8 **Dhanuthai K**, Rojanawatsirivej S, Thosaporn W, Kintarak S, Subarnbhesaj A, Darling M, Kryshtalskyj E, Chiang CP, Shin HI, Choi SY, Lee SS, Aminishakib P. Oral cancer: A multicenter study. *Med Oral Patol Oral Cir Bucal* 2018; **23**: e23-e29 [PMID: 29274153 DOI: 10.4317/medoral.21999]

9 **Parkin DM**, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin* 2005; **55**: 74-108 [PMID: 15761078 DOI: 10.3322/canjclin.55.2.74]

10 **Warnakulasuriya S**. Global epidemiology of oral and oropharyngeal cancer. *Oral Oncol* 2009; **45**: 309-316 [PMID: 18804401 DOI: 10.1016/j.oraloncology.2008.06.002]

11 **Osazuwa-Peters N**, Adjei Boakye E, Hussaini AS, Sujijantarat N, Ganesh RN, Snider M, Thompson D, Varvares MA. Characteristics and predictors of oral cancer knowledge in a predominantly African American community. *PLoS One* 2017; **12**: e0177787 [PMID: 28545057 DOI: 10.1371/journal.pone.0177787]

12 **Ogden GR**. Factors Affecting Survival for Oral Cancer. In: Warnakulasuriya S, Greenspan JS. Textbook of Oral Cancer. Textbooks in Contemporary Dentistry. Cham: Springer, 2020: 327-342 [DOI: 10.1007/978-3-030-32316-5_25]

13 **Sargeran K**, Murtomaa H, Safavi SMR, Vehkalahti MM, Teronen O. Survival after diagnosis of cancer of the oral cavity. *Br J Oral Maxillofac Surg* 2008; **46**: 187-191 [PMID: 18096283 DOI: 10.1016/j.bjoms.2007.11.004]

14 **National Cancer Institute: Surveillance, Epidemiology, and End Results Program**. SEER*Stat Databases: November 2018 Submission, Dictionary of SEER*Stat Variables. 2018. Available from: https://seer.cancer.gov/data-software/documentation/seerstat/nov2018/seerstat-variable-dictionary-nov2018.pdf

15 **Shiboski CH**, Schmidt BL, Jordan RC. Racial disparity in stage at diagnosis and survival among adults with oral cancer in the US. *Community Dent Oral Epidemiol* 2007; **35**: 233-240 [PMID: 17518970 DOI: 10.1111/j.0301-5661.2007.00334.x]

16 **Alcantud JCR**, Varela G, Santos-Buitrago B, Santos-García G, Jiménez MF. Analysis of survival for lung cancer resections cases with fuzzy and soft set theory in surgical decision making. *PLoS One* 2019; **14**: e0218283 [PMID: 31216304 DOI: 10.1371/journal.pone.0218283]

17 **National Cancer Institute: Surveillance, Epidemiology, and End Results Program**. About the SEER Registries. 2020. Available from: https://seer.cancer.gov/registries/

18 **National Cancer Institute: Surveillance, Epidemiology, and End Results Program**. ICD-O-3 Site Codes. 2020. Available from: https://training.seer.cancer.gov/head-neck/abstract-code-stage/codes.html

19 **National Cancer Institute: Surveillance, Epidemiology, and End Results Program**. SEER Combined/AJCC Cancer Staging. 2020. Available from: http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage

20 **National Cancer Institute: Surveillance, Epidemiology, and End Results Program**. Site Recode ICD-O-3/WHO 2008 Definition. 2020. Available from: http://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html

21 **National Cancer Institute: Surveillance, Epidemiology, and End Results Program**. Months Survived Based on Complete Dates, Calculation of Survival Time Fields. 2013. Available from: https://seer.cancer.gov/survivaltime/SurvivalTimeCalculation.pdf

22 **National Cancer Institute: Surveillance, Epidemiology, and End Results Program**. Months Survived Based on Complete Dates. 2013. Available from: https://seer.cancer.gov/survivaltime/

23 **Tshering Vogel DW**, Zbaeren P, Thoeny HC. Cancer of the oral cavity and oropharynx. *Cancer Imaging* 2010; **10**: 62-72 [PMID: 20233682 DOI: 10.1102/1470-7330.2010.0008]

24 **National Cancer Institute: Surveillance, Epidemiology, and End Results Program**. Cancer Stat Facts: Laryngeal Cancer. 2020. Available from: https://seer.cancer.gov/statfacts/html/laryn.html

25 **Gonzalez-Moles MA**, Esteban F, Rodriguez-Archilla A, Ruiz-Avila I, Gonzalez-Moles S. Importance of tumour thickness measurement in prognosis of tongue cancer. *Oral Oncol* 2002; **38**: 394-397 [PMID: 12076706 DOI: 10.1016/s1368-8375(01)00081-1]

26 **Yuen AP**, Lam KY, Wei WI, Lam KY, Ho CM, Chow TL, Yuen WF. A comparison of the prognostic significance of tumor diameter, length, width, thickness, area, volume, and clinicopathological features of oral tongue carcinoma. *Am J Surg* 2000; **180**: 139-143 [PMID: 11044531 DOI: 10.1016/s0002-9610(00)00433-5]

27 **Goldenberg D**, Brooksby C, Hollenbeak CS. Age as a determinant of outcomes for patients with oral cancer. *Oral Oncol* 2009; **45**: e57-e61 [PMID: 19362043 DOI: 10.1016/j.oraloncology.2009.01.011]

28 **Davidson BJ**, Root WA, Trock BJ. Age and survival from squamous cell carcinoma of the oral tongue. *Head Neck* 2001; **23**: 273-279 [PMID: 11400227 DOI: 10.1002/hed.1030]

Published by **Baishideng Publishing Group Inc**
7041 Koll Center Parkway, Suite 160, Pleasanton, CA 94566, USA
**Telephone:** +1-925-3991568
**E-mail:** bpgoffice@wjgnet.com
**Help Desk:** https://www.f6publishing.com/helpdesk
https://www.wjgnet.com