UNIVERSITY OF CALIFORNIA,
IRVINE

Determinants of Health in Oral Cancer Patients by Treatment and Transcriptome

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Epidemiology

by

John Paul Schomberg

Dissertation Committee:
Professor Hoda Anton-Culver, Chair
Associate Professor Argyrios Ziogas
Assistant Professor Trina Norden-Krichmar

2018

DEDICATION

To

my family

Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.

Madame Curie

Those who have learned to walk on the threshold of the unknown worlds, by means of what are commonly termed par excellence the exact sciences, may then, with the fair white wings of imagination, hope to soar further into the unexplored amidst which we live.

Ada Lovelace

Truth is not what you want it to be; it is what it is, and you must bend to its power or live a lie.

Miyamoto Musashi

No individual has any right to come into the world and go out of it without leaving behind him distinct and legitimate reasons for having passed through it.

George Washington Carver

All truths are easy to understand one they are discovered; the point is to discover them.

Galileo Galilei

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my committee chair, Hoda Anton-Culver, who provided academic support throughout the entirety of this dissertation composition. Dr. Anton-Culver's unwavering dedication to the success of her students was invaluable in seeing this dissertation to its completion. Working with Dr. Anton-Culver has made provided me with the ability to set goals with clarity and focus. I am forever indebted to Dr. Anton-Culver for the time and faith she has invested in my future as an epidemiologist.

I would like to thank my committee members, Professor Argyrios Ziogas, and Professor Trina Norden-Krichmar, for the many hours spent in revising manuscripts and providing much needed feedback as I progressed along the path of academic distinction.

In addition, special thanks are extended to the head and neck cancer patients treated at the University of California Irvine Chao Cancer Center. The spirit and resilience of these patients served as an enduring source of inspiration in the creation of this dissertation.

CURRICULUM VITAE

John Paul Schomberg

2003 B.S. in Dietetics/Nutrition, Iowa State University, Ames Iowa

2011 Master's of Public Health University of California Los Angeles

2013-2018 Graduate Student Researcher Department of Epidemiology, School of Medicine. University of California Irvine

2018 Ph.D. in Epidemiology
University of California, Irvine

FIELD OF STUDY

Determinants of Treatment Response in Oral Cancer

PUBLICATIONS

- Pediatr Nephrol. Sevelamer Carbonate Increases Serum Bicarbonate in Pediatric Dialysis Patients. 2010 Feb. Gonzalez, E, Schomberg J, Amin N, Salusky IB, Zaritsky J.
- Journal of Renal Nutrition Journal of Renal Nutrition 2015 Oct Successful Conversion from Parenteral Paricalcitol to Pulse Oral Calcitriol for the Management of Secondary Hyperparathyroidism in Hemodialysis K. Kalantar-Zadeh
J. Kumar; Gia Tran; J. Schomberg; E. Streja; M. Pahl (In Peer Review)
- PLOS ONE Supplementing Public Health Inspection via Social Media 2016 Oct J. Schomberg, O. Haimson, G. Hayes, H. Anton-Culver
- Oral Oncology Identification of a Gene Expression Signature Predicting Survival in Oral Cavity Squamous Cell Carcinoma Using Monte Carlo Cross Validation 2017 Feb. J. Schomberg, A. Ziogas, H. Anton-Culver, T. Norden-Krichmar.

ABSTRACT OF THE DISSERTATION

Determinants of Health in Oral Cancer Patients

By

John Paul Schomberg

Doctor of Philosophy in Epidemiology

University of California, Irvine, 2018

Professor Hoda Anton-Culver, Chair

Oral cancer is a disease related to multiple factors. In the United States, approximately 50,000 people are diagnosed with cancer of the head and neck each year. Unfortunately, even though there have been some improvements in care, 10,000 head and neck cancer patients die each year. The focus of this dissertation is to determine those patients that would respond best to treatment by following quality of care guidelines, utilizing genetic signatures, and identifying genetic targets for treatment through machine learning analysis.

This dissertation begins by specifically addressing the differences in survival of patients that meet and follow the National Comprehensive Cancer Network (NCCN) guidelines for the recommendation of chemotherapy. This study is a well-powered analysis of 37,985 patients selected from the California cancer registry. It was found that patients have significantly improved survival when their provider prescribes chemotherapy as recommended by NCCN.

In Chapter 3, gene expression signatures were utilized to predict patient response to treatment. An aggregate signature was identified using a high dimensional dataset with a relatively low number of patient samples (n=257). By permuting the dataset 100 times via Monte Carlo cross validation and then performing differential expression analysis between treatment responders and non-responders within each permuted dataset, this study was able to identify genes that were differentially expressed across multiple permutations and utilize those gene expression values within a final aggregated signature predicting treatment response.

Chapter 4 utilizes the same gene expression data set in a different way by applying a machine learning method known as random forest to rank influential genes and evaluate the pathways within which those genes reside. Integrated with this machine learning analysis is the application of chemical informatics to identify those small molecules in an FDA-approved drug database and a Traditional Chinese Medicine database that meet similarity criteria when measured against a reference ligand known to bind to a drug target site.

This dissertation advances the knowledge of effective treatments in oral cancer, and provides greater understanding of the genetic pathways influencing treatment response.

CHAPTER 1

INTRODUCTION

Epidemiology of Head and Neck Cancer

Head and neck cancers can be defined as cancers in the upper airway and/or digestive tract found in oral cavity, laryngeal, pharyngeal, and oropharyngeal, and hypo-pharyngeal tissues. Such cancers make up 3% of cancers diagnosed each year[1,2]. Head and neck cancer incidence has declined from 25 cases per 100,000 at risk in the 1990s to 15 cases per 100,000 at risk in the present day[3]. While the decrease in head and neck cancer incidence may be due to a drop in tobacco use [4,5], the mortality associated with oral cavity cancers has not changed significantly in the last twenty years, with the exception of African American patients in whom we have seen a significant decrease in mortality associated with the disease[6]. While Human Papilloma Virus (HPV) positive patients have been observed to have an improved survival and response to treatment when compared to HPV negative patients these patients still make up the minority of head and neck cancers[7]. Thus, the decline in mortality could be attributed to decrease in smoking, increases in HPV positive cases, or a reduction in healthcare disparity for African-Americans. The following studies will focus upon the determinants of treatment response in a subset of head and neck cancer patients, oral cancer patients. Specifically, this study will measure the effect of NCCN guideline adherence on oral cancer patient survival, the identification of a gene expression signature that predicts oral cancer patient treatment response, and the use of machine learning methods to identify genes and gene pathways influencing treatment response. These studies will improve understanding of the determinants of survival in

oral cancer patients while simultaneously outlining methods that can be utilized for identifying genetic signatures and gene targets in rare cancer.

Human Papilloma Virus and Oral Cancer Etiology

Human Papilloma Virus has been shown to be a driver of 99% of cervical cancers[8]. This virus is also associated with development of oral cancer in men and women[9,10]. Male patients with oral cancer are reported to have twice the prevalence of testing sero-positive for human papilloma virus[11–13]. However, not all patients testing seropositive for HPV have oral tumors that are driven by HPV[14–17]. There are distinct molecular signatures associated with oral cancers associated with alcohol and tobacco use, and there are signatures associated with HPV driven tumors. Currently these signatures are based upon mutations in TP53, CDKNA, EGFR, and PIK3A[18]. HPV p18, or p16 are the only molecular signatures that have been shown to be predictive and prognostic[19–21]. These signatures are used to guide treatment of oropharyngeal tumors specifically. HPV expressed proteins E6 and E7 specifically suppress TP53 which confers cell immortality. Fortunately HPV genome expression is not associated with the mutations(MYC CCND1, SOX2) in tobacco driven tumors which are associated with treatment resistance[17]. It is for this reason that HPV sero-status is a useful predictor of treatment response.

Role of Alcohol and Tobacco in Oral Cancer

Alcohol and tobacco are two known risk factors for oral cancer. Alcohol alone has been shown to account for a small proportion of the attributable risk[22]. However, alcohol and tobacco use in combination have shown to have a multiplicative effect on risk[22]. Alcohol is believed to multiply the cancer risk associated with tobacco use by dehydrating cell walls and increasing their permeability to carcinogens present in tobacco products. Chronic alcohol use can lead to antioxidant deficiencies which predispose patients to the oxidative damage and DNA single strand breaks which lead to mutations resulting in oral cancer. Tobacco products are known to carry several carcinogens, and the ignition of tobacco during tobacco produces polycyclic hydrocarbons that can also produce oxidative damage on cell DNA[23–26]. Alcohol and tobacco use in oral cancer has changed by time period in a way that is gender dependent[27]. Evaluation of response to treatment by period, gender, and tumor type will be an important part of evaluating persistence of gender disparity across combination of clinical features.

Differences in Exposure by Gender

In Oral cavity cancers environmental exposures like alcohol and tobacco use are well established in their association with head and neck cancer, as is the greater risk to male patients[28]. A 2013 study of SEER data showed that for the past 20 years the rate of head and neck cancers in men have been double that in women. The relationship is consistent across race, and ethnicity. The proportion of cases associated with HPV infection is twice as high in men as it is in women. This study reported that 10-15% of

cases of head and neck cancer in women could not be attributed to alcohol or smoking exposures. The California Cancer Registry is a database of cancer cases following SEER reporting guidelines. The strength of such a database is the large number of cases tracked between the years 1988 and 2012 and the high number of variables tracked in this registry that can inform and improve upon analyses by adjusting for confounding and biases. This database will be used to assess the persistence of treatment response disparities between men and women. Findings generated by this study will further illuminate the discussion of possible mechanisms to which oral cavity cancers can be attributed. This study will strengthen the argument for or against the attribution of gender disparities to statistical artifact.

In a 2009 pooled analysis of head and neck cancer case control studies [29] authors stated there was a 33% increased population attributable risk (PAR) of HNSCC for tobacco users, and another 35% increased PAR for patients reporting both alcohol and tobacco use, finally there was a 4% increased PAR for those who reported alcohol intake alone. Women were reported to have a lower PAR than for men. It is also important to note that while men had a higher PAR for alcohol and tobacco than for tobacco alone women did not. These results indicate that HNSCC cases in women cannot be attributed to alcohol and tobacco use to the same extent they are attributed to alcohol and tobacco exposure in men. This study reports that 42% of cases in women cannot be attributed to alcohol or tobacco exposure. A proportion of these cases could be attributed to human papilloma virus (HPV) infection however it is unlikely that all 42% of female cases not attributed to alcohol or tobacco exposures can be attributed to HPV

infection alone. HPV infection is known to infect a greater proportion of male than female patients, it has been reported that male HNSCC cases that cannot be attributed to tobacco or alcohol account for only 26% of cases. If we assume that all cases that cannot be attributed to alcohol and tobacco can be attributed to HPV in men, and that women can attribute a similar proportion of cases to HPV then there would still be 16% of cases in women that cannot be attributed to HPV. However this number is likely to be higher as studies examining the distribution of HPV positive sero-status across the HNSCC patient population all report that the prevalence of HPV positive sero-status is greater in males than females.

The Role of Race and Socioeconomic Status in Oral Cancer Survival

African-American patients have been shown to have lower risk of survival when compared to white patients. These differences in survival have been attributed to different distributions in the size of patient clinical stage, tumor size and tumor grade. There is also an association between race and socioeconomic status with African-American. Socioeconomic status is also predictive of patient access to all treatment modalities. Other researchers have shown an association between socio economic status and access to chemotherapy. It is important to take steps to adjust predictive models for socioeconomic status before assessing the disparities in access to care between racial groups. Of course the power to detect such differences is also dependent upon well powered representative samples such as those using extracted from cancer registries. Some believe that disparities in survival by race can be attributed to poor access to early screening that would provide access to care when oral

cancers can be treated with a greater rate of success. Others have made the point that mistrust of medical institutions may be significantly different when stratified by race and thus mistrust may be an additional barrier to access to recommended treatment. Finally, it is possible that disparities in treatment response by race can be attributed in part to differences in those genes that metabolize chemotherapies. Cytochrome P450 genes within xenobiotic pathway are distributed differently by race and thus may play a role in disparities. For a variety of reasons experimental trials used to verify the effect of chemotherapies are not racially diverse. Using homogeneous populations when testing new therapeutics increases the ability of researchers to detect a treatment effect, however restrictions on racial disparity also limit the exportability of trial findings to a variety of racial groups. Validation of chemotherapeutics effectiveness when administered according to recommended guidelines to ethnically diverse sample of patients would serve as an appropriate first step in determining the effectiveness of such therapies across racial and socioeconomic groups.

Molecular Signatures in Oral Cancer

There are few studies that have identified a group of genes for predicting those HPV-negative HNSCC patients that will receive greater benefit from high intensity radiation treatment. To date the most widely used molecular signature guiding HNSCC treatment is HPV status. However, HPV infection preferentially infects oropharyngeal tissues which make up only 15% of HNSCC[30]. Pilot studies examining the role of genetic markers in head and neck cancer have reported promising but mixed results, although these studies also reported limitations in interpretation due to small sample size[31–33].

Past studies have focused upon gene expression and mutation in known oncogenes: Epidermal Growth Factor Receptor (*EGFR),* Transforming Growth Factor Beta *(TGFRB),* Kirsten Rat Sarcoma Viral Oncogene Homolog (KRAS), and Tumor Protein 53 (*TP53)*. However, those results have been difficult to interpret due to limitations of the study to mutation status or limitations of the study to the expression of single gene. Other studies have found that canonical oncogenes like EGFR are regulators of cell growth, migration and survival, and that EGFR is overexpressed in 80 out of 100 head and neck cancers. Mutations and epigenetic changes in these classical oncogenes have been associated with proliferation, apoptosis, inflammation, angiogenesis and DNA repair in head and neck cancers.

Other studies report using microRNA expression to predict overall survival across laryngeal, pharyngeal , and oral tissues[34]. There have been multiple studies that have identified genetic markers that improve prediction of overall survival when HPV status is known[35–38]. Unfortunately, there has been little focus in squamous cell carcinoma in the oral cavity, a HNSCC subgroup that is known to respond significantly worse to treatment than patients with Oropharyngeal Squamous Cell Carcinoma (OPSCC). It is also important to note that the authors of the Wong et al[34] study found that there was significant genetic heterogeneity in each HNSCC tissue subtype. Therefore, this study focuses solely upon treatment sensitivity in OSCC specifically. This study contributes evidence that even in the high risk populations of head and neck cancer patients, such as tobacco users and alcohol abusers, [39] patients can be stratified into meaningful treatment groups based upon measures of gene expression in oral cavity tumor cells.

New Chemotherapy Treatments in Oral Cancer

In the last ten years one new medication has been approved for the treatment of head and neck cancer patients. The most recent drug (PD-1 inhibitor Pembrolizumab) was previously approved for patients with stomach cancer. Taxol based medications Erbitux and Paclitaxel have been found to induce greater radiation sensitivity in patients with cancer of the head and neck. The standard or care in oral cancer is surgery, and radiation therapy. Patients with tumors greater than clinical stage three, and node positive disease meet criteria for systemic/chemotherapy. When identifying cancer targets in rarer cancers such as oral cancer, researchers must rely upon datasets implicitly smaller in size. Several researchers have used machine learning methods to identify signals within genetic data. Permuting genetic datasets with the machine learning method known as random forest is one method that has been shown to be effective in classifying patients via high dimensional genomic data.

Chemical informatics is a widely utilized first step when identifying which ligands that should be considered in high throughput screening. Using the ranked list of genes generated from random forest can be purposed towards the goal of using ligands known to bind proteins expressed by influential genes in virtual screening applications. A single pipeline partnering bioinformatics and chemical informatics has the potential to improve interface and utility of drug discovery pipeline. The final step would be to validate such a process against existing literature to verify the effectiveness of such applications.

Oral cancer is a disease that occurs less often than breast, lung, or prostate cancer. Understanding the effectiveness of the standard of care recommended by the NCCN, and the equity of the distribution of that care is important in that contributes to knowledge on the efficacy of such treatment quality guidelines, and it highlights sources of disparity in treatment in oral cancer patients. Specialized methodologies should be used when evaluating the genetic information for a small sample of patients. This work identifies how Monte Carlo cross validation can be used to strengthen the signal of genes, and to form an aggregated signature that predicts treatment response. This work not only generates signatures that could be utilized within a clinical setting but offers insight to those pathways playing an important role in survival in head and neck cancer. This work is furthered through the application of random forest a machine learning method that performs well with high dimensional datasets with low numbers of observations. By integrating this approach with network analysis and chemical informatics, this work identified drug leads that actively inhibit progression of oral cancer cells. This activity was identified via literature review of all drug leads with greater than 50% Tanimoto similarity. In addition, this work identifies influential genes and pathways in oral cancer specifically in those patients treated with platinum based chemotherapy and in non-platinum based chemotherapy.

Chapter 2

NCCN Guideline Adherence and Oral Cancer Treatment and Response

Introduction

Head and neck cancers can be defined as cancers in the upper airway and/or digestive tract found in oral cavity, laryngeal, pharyngeal, oropharyngeal, and hypo-pharyngeal tissues. Such cancers make up 3% of cancers diagnosed each year[1,2]. Head and neck cancer incidence has declined from 25 cases per 100,000 at risk in the 1990s to 15 cases per 100,000 at risk in the present day[3]. The decrease in head and neck cancer incidence has made progress, (521 per 100,000 at risk to 435 per 100,000 at risk annually) [4,5] from 1988-2012. Additionally, the mortality associated with oral cavity cancers has decreased over the last twenty years (251 deaths per 100000 to 161 deaths per 100000 annually)[4,5]. African American patients have also made progress in that there has been a decrease in mortality associated with oral cancer in this group of patients [6]. However, most recent reports still show that racial disparities persist between African-American and White patients with oral cancer with 65% of White patients reporting 5 year survival vs. 48% in African-Americans[7–12]. It is important to note that there have been few studies that closely examine whether racial disparities in oral cancer treatment response are associated with disproportionate levels of adherence to quality of care guidelines like the National Comprehensive Cancer Network (NCCN) recommendations. There is a need to better understand the benefits/detriments associated with NCCN guideline adherence for providers treating patients with oral cancer. Evaluating the clinical and demographic variables associated

with NCCN guideline adherence will improve understanding of how access to quality care may be improved.

The NCCN is a non-profit organization of 27 cancer centers across the world dedicated to “improving the effectiveness, and efficiency of cancer care so that patients can live better lives”[13]. The NCCN provides guidelines on the treatment of common and rare cancers and offers annually updated treatment algorithms reflecting the highest standards of patient care. Oral cancer treatment guidelines provided by NCCN offer guidance on the appropriate implementation of chemotherapy for each oral cancer patient. One aim of this study will be to identify if NCCN guidelines are applied uniformly across racial groups, and if all racial groups respond without significant differences in overall survival when NCCN guidelines are applied uniformly across racial groups. While there have been few studies that report upon racial disparities in oral cancer patients there are many studies that report upon disparities in cancer treatment by race for other anatomic sites[14–17]. African-American patients have a 33% greater chance of dying of all-site cancer compared to White patients[18]. This disparity has been attributed to the observation that African-American patients are more likely to be diagnosed with regional, or metastatic tumors than White patients[19,20]. Even after adjusting for stage of cancer some studies have continued to note significant disparities in overall survival between White and African-American patients[7–12,21], indicating that stage of disease alone may not be the only risk factor for poor survival.

Treatment response is predicted in large part by the grade and size of a tumor, the degree of nodal invasion, and the type of treatment which is applied to a patient. It is equally true that the mode of treatment (surgery, chemotherapy), is based upon the anatomic site, grade, nodal status, and tumor size of a given patient. The 2017 NCCN treatment algorithm recommends systemic/chemotherapy to any patient with node positive disease diagnosed as Clinical Stage III or greater Racial disparities in cancer treatment response have been attributed to differences in the distribution of advanced disease due to lack of appropriate screening. However, if the significance of race as a predictor of NCCN recommendation adherence by providers persists across combinations of confounding variables then it strengthens the argument that differences in overall survival by race may be attributed to variables other than advanced stage of disease due to lack of screening.

Disparities in overall survival between African-American and White patients have been attributed to quality of care measures such as adherence of physician recommendation to cancer care guidelines. Delivery of quality healthcare guided by physician recommendation often is dependent upon a group of best practices or standards that can serve to guide treatment. Measurement of racial disparity can only be assessed after confounding variables like social status, age, gender, tumor grade, and treatment type have been accounted for. Oral cancer is a rare cancer in which patient race has been shown to be a strong predictor of overall survival. Analysis of data rich resources such as the California Cancer Registry can be performed to identify explanatory variables (such as physician adherence to NCCN chemotherapy assignment guidelines)

that may further the understanding of mechanisms of observed racial disparities in oral cancer that have not yet been explored.

This study will specifically aim to address:

1) The effect of NCCN guideline adherence on survival in those patients recommended to receive chemotherapy by NCCN (Stage III- Stage IV node positive)
2) Whether racial disparity in overall survival persists in patients that are provided NCCN compliant care.
3) Which factors (including race) drive provider assigned chemotherapy treatment when chemotherapy is not recommended by NCCN (Clinical Stage I-Clinical Stage III Node Negative)?
4) What is the survival benefit associated with chemotherapy use in those patients not recommended by NCCN to receive chemotherapy (Clinical Stage I- Clinical Stage III Node Negative)?

Methods

Descriptive Study of Population

37,985 oral cavity cancer patients were selected from the California Cancer Registry (CCR) from 1988 to 2012 after exclusion of cases with missing variables. “Oral cavity cancers” were inclusive of all patients within the CCR with cancer of the tongue, lip, gum, alveolar ridge, floor of mouth, and inner cheek. 8 oral cancer patients were not included in analyses as they were diagnosed at autopsy and so received no treatment. 25 patients were diagnosed via death records and thus were also excluded from study. Only patients with malignant cancers within the CCR were included in this study, 1883 benign tumors were excluded. 10789 patients without tumor grade treatment were also excluded from analyses. 2613 patients without nodal status were not included in the analyses of this study. Also, 5281 patients with unknown tumor size were also excluded from analyses. Finally, 458 patients with unknown chemotherapy treatment were also excluded. There were 1955 patients with missing information on surgery treatment, these patients were included within analyses of all patients as they did have information on chemotherapy treatment and other clinical and demographic features thus improving the ability of this study to adjust for such confounders. This left 37,985 patients remaining for this analysis. When patients were grouped by NCCN recommendation status (a variable created through the combination of clinical stage and nodal status) there were 2416 patients that were without NCCN recommendation due to either lack of nodal status or lack of information on clinical stage. Remaining were 23,521 patients not recommended to receive chemotherapy by NCCN (clinical stage I-III and node

negative), and 12,048 patients recommended to receive chemotherapy treatment by NCCN (clinical stage III node positive). (Supplemental Figure 1, Supplemental Figure 2)

A goal of this study was to specifically examine if adherence to NCCN recommendations significantly improved the quality of patient care (measured in overall survival). Treatment response was measured in overall survival in months. Receiving chemotherapy was defined as receiving a chemotherapy treatment post diagnosis for any duration. The standard of care according to NCCN guidelines for treatment of head and neck cancers calls for the use of chemotherapy for clinical stage III node positive patients and all higher stages in combination with surgery and or radiation. Radiation treatment adjusted for in this analysis refers to beam radiation therapy. While surgery refers to any surgical procedure performed for the purpose of treating the cancer in question. Racial groups included in this analysis were White, African-American, Asian, Hispanic, and “Other”. “Other” race consists of a combination of all races not aligning within the first four groups. Native American patients were included within the “other” group as small sample sizes prevented stable estimates of effect size. Tumor size <25mm is a binary measure using the greatest tumor dimension of all dimensions measured.

Statistical Methods

This is a retrospective, population-based, case-only study of squamous cell oral cancer reported to California Cancer Registry between January 1st, 1988 and December 31st, 2012. California Cancer Registry case reporting is estimated to be 99% for the entire state of California, with follow-up completion rates exceeding 95%. Case selection criteria included those patients 18 years of age or older and had to be the first cancer diagnosed (Fig. 1). Age at diagnosis was treated either as a continuous variable or as a categorical variable with four groups (younger than 45 years, 45–54 years, 55–69 years, and 70 years or older). Tumor characteristic included Clinical Stage, tumor grade, Tumor Size and Nodal Status. The last date of patient contact used was either the last date of contact listed or the date of death. Adherence to National Comprehensive Cancer Network recommended therapy was chosen as a measure of the quality of cancer care received and considered as a gold standard that most oral cavity cancer patients should be allowed. Provider adherence with treatment recommendations for oral cavity cancer was based on NCCN recommendations for surgery and chemotherapy treatment. According to the most recent recommendations (2017) for Stages I–III (node negative) patients, surgical treatment and or radiation therapies were considered in compliance with NCCN recommendations. NCCN does not provide recommendation for the use of chemotherapy in oral cancer patients with less than stage III node negative cancer. For cases of stages III-IV (node positive) oral cancer, surgical treatment and radiation and chemotherapy are recommended. Number of Chemotherapy agents and specific type of beam radiation were not considered for this study nor were the approaches of surgical intervention. This study defined NCCN

guideline adherence with chemotherapy recommendation as those individuals receiving chemotherapy as their primary treatment or in conjunction with another treatment modality (surgery, radiation) that met NCCN requirements for recommendation of systemic/chemotherapy (Node positive disease of Clinical Stage III or greater). Socioeconomic status data in the California Cancer Registry represent a composite of measures of the census tract in which the patient resides at diagnosis (e.g., education, income, cost of living, and occupation type). This variable was categorized from lowest to highest quintile. This representation of SES has been used in past epidemiologic journal publications on various types of cancers. NCCN compliance was considered across the sample subset by modality combinations (radiation, surgery, radiation and surgery, no radiation and no surgery).

Survival Analysis

Cox regression models predicting treatment response in overall survival (measured in months) were also produced. Cox models adjusted for tumor size, nodal status, surgery, grade, socioeconomic quintile, age, gender, and race. Tumor size measured in cubic mm was converted into a binary measure of greater than 26 cubic mm or less than or equal to 26 cubic mm. Multivariable logistic regression analysis was conducted to measure the probability of noncompliance with NCCN recommendations for chemotherapy use. The second measure of interest was disease-specific survival. Survival analysis was performed using the Kaplan-Meier estimate of survival probability and log-rank test to visualize difference in survival according to NCCN recommendation for chemotherapy (Figure 2, Figure 3). After verifying the proportional hazards

assumption, a Cox proportional hazards model was used to estimate the effect of each variable on overall survival, with clinical stage treated as a strata and not a predictor. The model compared post diagnosis survival time in patients with the same clinical stage of disease and produced a coefficient weighted by stage for other factors in the model. Hazard ratios (HR) and 95% confidence intervals (CI) were produced by these analyses. All statistical analysis was conducted using SAS 9.4.

Results:

Descriptive Results:

This study identified 37985 patients from the CCR who met criteria for complete clinical and treatment modality information. The median age at diagnosis was 63 years old with a standard deviation of $\pm$13.7 years. 12048 of 37985 (31.7%) of patients met criteria for NCCN recommendation of chemotherapy. Of those who were recommended to receive chemotherapy 4499 of 12048 (33.8%) were compliant with NCCN recommendations. Of those 23521 patients with early stage node negative disease, 1618 of 23521 (26.4%) were still prescribed chemotherapy by providers (Supplemental Table 1). A summary of descriptive analyses can be found in Table 1. When subdivided by anatomic site it can be seen that patients with cancer of the tongue were more likely to receive chemotherapy when cancers were node negative and in early stage. 11% of early stage tongue cancer patients received chemotherapy compared with 6.2% and 6.3% in floor of mouth and gum cancers, and .7% in gum cancers. In patients with “other oral cavity cancers the proportion receiving treatment in contrast with NCCN recommendation was 23%. Frequency of patients receiving treatment by stage, nodal

status, and site can be viewed in Table 2.7. When chemotherapy was not recommended by NCCN, greater proportions of African-American, Hispanic, Asian, patients received chemotherapy (11.1%, 8.3%, 7.1%) respectively than the proportion of white patients (6.5%) receiving chemotherapy that contradicted NCCN recommendations. Patients with missing surgical treatment status were included in this analyses to provide additional power in detecting the effect of NCCN compliance on overall survival in those patients recommended to receive chemotherapy by NCCN and those that did not.

Results of Multivariable Regression Predicting NCCN Compliance

A multivariable logistic regression model was produced to measure the effect of clinical stage, weighted clinical and demographic features on the probability of provider adherence to NCCN treatment guidelines for oral cavity cancers (Table 2). Clinical stage strata were used to better assess the contribution of each variable to model prediction. Demographic and clinical feature types were associated with compliance with NCCN recommendation for chemotherapy (Table 2). Patients of lower socio economic status, were more likely to have providers that did not adhere to NCCN recommendation for chemotherapy in comparison to those patients in the highest quintile of socioeconomic status (OR 1.54, 95% CI 1.2-1.95, OR 1.38 95% CI 1.1-1.7, OR 1.24, 95% CI 1-1.5, OR 1.26, 95% CI 1-1.5.) for the first, second, third and fourth SES quintiles respectively. Younger patients in the first four age groups were less likely to have providers that do not adhere to NCCN recommendations (OR .34, 95% CI .37-.59, OR .34 95% CI .27-.41, OR .47, 95% CI .4-.55) for those patients younger than 45,

45-54, and 55-69 respectively. There was no statistically significant difference between any racial group and white patients in terms of their likelihood to receive chemotherapy. Clinical features like tumor grade III (OR 1.25 95% CI 1.0-1.4), and tumor size <26 mm (OR 1.44, 1.2-1.6) were all more likely to be predictive of noncompliance with NCCN recommendations for chemotherapy (Table 2).

When considering those clinical and demographic features that were predictive of the use of chemotherapy in those patients with early stage node negative disease, this study found that tumor grade and tumor size were significant predictors of whether a patient would receive chemotherapy to treat early stage disease (contradicting NCCN criteria). Patients with Tumor size < 26 mm had an increased odds (OR 2.1, 95% CI 1.4-3.0) that chemotherapy would not be given. Patient with Tumor grade G1-G2 had an increased odds of having chemotherapy withheld when compared to G3 patients (OR 1.9 95% CI 1.5-2.3). There was no significant difference between White and African-American, Hispanic, or Asian patients in their likelihood to receive chemotherapy treatment that contradicted NCCN guidelines (Table 3).

Survival Analysis Results

Kaplan Meier survival analysis showed a statistically significant difference between patients receiving care that was compliant, and noncompliant with NCCN guidelines recommending use of chemotherapy as part of oral cancer treatment in node negative Stage I-III (Figure2) and node negative Stage III-IV (Figure 3) respectively. The survival of patients in the node negative low stage group were significantly different from one another (logrank p-value <.0001), with those patients in compliance with NCCN

recommendations (not receiving chemotherapy) having better survival overall. This was evident in that the median survival for those patients receiving chemotherapy was 36 months, while median survival in those patients not receiving chemotherapy was 81 months. The survival of patients in the node positive late clinical stage group was significantly different between patients who received chemotherapy and those who did not (p-value<.0001). Patients receiving chemotherapy with late stage node positive tumors had significantly better survival compared to those patients that did not receive chemotherapy. Median survival in the node positive high stage groups was 21 months in those patients not receiving chemotherapy, and 32 months in those patients receiving chemotherapy.

A multivariable Cox regression model was used to measure the effect of chemotherapy treatment in patient groups stratified by clinical stage and nodal status. Cox models weighted clinical and demographic features on the probability of survival in patients meeting requirements for NCCN recommendation of chemotherapy. Compliance with NCCN recommendations was significantly associated with survival in those patients who also received surgery and radiation. Compliance with chemotherapy use as recommended by NCCN, did not have a significant effect in those patients who received radiation alone, surgery alone, or no surgery and no radiation. All patients included in analysis had information on clinical stage and nodal status and thus were eligible to be assigned NCCN recommendation on whether they should be receive chemotherapy treatment. Multivariable Cox regression analysis confirmed the known negative effects of older age, advanced tumor grade (Table 4). After controlling for additional

confounding factors, compliance with NCCN recommendations for chemotherapy was statistically significant in predicting improved survival in oral cancer patients. Patients receiving compliant care had a 33% decrease in the risk of death (HR .67, 95% CI .6-.8). In contrast, patients that did not meet the criteria for NCCN recommendation of chemotherapy (Stage I-III node negative) had a 50% increased odds of death (HR 1.5, 1.4-1.7) when chemotherapy was received. Patients receiving radiation alone, or no surgery and no radiation all had increased odds of death when chemotherapy was received (HR 1.4, 95% CI 1.2-1.7), (HR 1.9, 95% CI 1.5-2.3) for radiation alone, or no surgery and no radiation respectively (Table 5).

Discussion:

Summary

It is desirable to adopt measurements of adherence with NCCN guidelines as a measure of quality in that it provides a simple means of comparing the quality of care across different patient groups. Use of NCCN guidelines also produces a benchmark against which future interventions can be compared and contrasted. Analyses described within this paper present a population based study that groups patients by specific treatment modality types listed in the CCR to identify the probability of access to NCCN adherent care across clinical and demographic groups. This study was able to measure the magnitude of effect of NCCN compliance across combinations of different treatment modalities. These analyses examined the effects not only in those patients where chemotherapy treatment is recommended (late stage node positive oral cancer) by NCCN but also where it was not (early stage node negative oral cancer). The effect of chemotherapy use on survival in those patients not recommended to receive chemotherapy by NCCN, and the factors associated with treatment assignment are important when considering the role of NCCN recommendations across all stages of disease, and the importance of providing guidance on restricting those treatments due to their association with increased mortality at specific stages of disease.

In those patients with late stage oral cancer, when providers were also adherent to care guidelines of recommendations for surgery and radiation treatment, there was no significant difference in provider use of chemotherapy between White patients and any other racial group (African-American, Asian, Hispanic, Other). Compliance with NCCN

recommendation for chemotherapy was less likely for African-Americans that received surgery alone as a treatment modality. It is possible that the higher rate of comorbidities in African-American[22–24] patients drove provider noncompliance due to greater health risk associated with a regimen of surgery and radiation and chemotherapy as recommended by NCCN. Comorbidities such as diabetes and hypertension have been shown to be risk factors for surgical complications[25–27]. Currently, the NCCN does not provide guidance on tailoring therapy by patient comorbidity. Such analysis is also not possible to perform with CCR data as there are no data related to comorbid status while receiving treatment. Further analyses with datasets tailored to address these issues are recommended. These results support adherence to NCCN recommendations across racial groups as a key step in the elimination of racial disparity in oral cancer treatment.

This study's strength is its use of a population based sample covering the entire state of California. This is the first study using a population of this size in determining the survival benefit associated with provider adherence to NCCN oral cancer guidelines. This is a dataset tracking the incidence of cancer prospectively from point of diagnosis that aims to track 100% of cancer cases in the state of California through multiple reporting mechanisms. This study was designed and powered to identify the magnitude of effect associated with adherence to NCCN recommended treatment. It is notable that the majority of patients observed in this study were of lower clinical stage (< Clinical Stage 4 and node negative). These patients are not recommended to receive chemotherapy by NCCN and yet 7% of those patients received chemotherapy

contradicting NCCN recommendations. Conversely, only 37% of patients that were recommended to receive chemotherapy by NCCN actually received the recommended treatment. The tendency for patients not to receive chemotherapy treatment could belie several mechanisms: chemotherapy access is gated by income (this is supported by the association with socioeconomic status), chemotherapy could also be gated by access to insurance providers offering full coverage of this treatment, chemotherapy access is not provided in advanced tumors due to impact on quality of life in patients with disease that is unlikely to respond to treatment, chemotherapy is declined by patients due to mistrust of medical institutions and is not reported as declination.

For those patients receiving chemotherapy treatment for early stage tumors tumor size was predictive of survival specifically in those patients that did not receive surgery. It is tempting to infer that if a patient does not receive surgery as recommended, and a tumor is large, then surgery may be withheld because a tumor is inoperable. In such a situation where the recommended treatment (surgery) cannot be provided, a provider may feel it is their duty to provide other therapies regardless of whether those therapies are recommended by NCCN. This inference is further supported by Table 7 which shows that the greatest proportion of patients receiving chemotherapy in contrast with NCCN guidelines are those with tongue and “other” oral cavity cancers. Surgeries requiring removal or damage of nervous tissue enervating the tongue are deemed inoperable by a majority of oral surgeons. It may be that providers that cannot offer surgical intervention feel that offering an unsupported therapy like chemotherapy is better than no intervention at all. This study provides evidence that providing

chemotherapy in such situations is associated with a detrimental effect on patient survival. Similar effects were identified in patients that had both radiation and surgery withheld. This study should provide caution to providers that wish to exhaust all possible therapies in patient treatment, a case must be made for patient quality of life during treatment especially when treatment is not associated with improved survival.

It should be noted that chemotherapy is not prescribed to extend life alone. For patients with terminal disease chemotherapy may be prescribed to reduce discomfort while the patient is receiving hospice/ end of life care. Data on the use of hospice services and the proportion of patients receiving chemotherapy while on hospice does not exist within the CCR, Further work is needed examining the palliative use of chemotherapy in oral cancer patients to determine if a disproportionate use of chemotherapy in early stage oral cancer on hospice contributes to the negative association between survival and chemotherapy use in early stage patients. This confounding would seem unlikely as hospice care is typically utilized in late stage patients with no treatment options remaining. It is only when early stage patients advance to later stages that NCCN recommends combining treatment modalities of surgery, radiation, and chemotherapy. As long as further treatments aimed to control disease remain, a patient would not be considered for hospice unless recommended care was declined. This study did control for the effect of care declination by excluding all patients declining chemotherapy care from analyses.

Lower SES was a predictor of a greater likelihood of not receiving chemotherapy regardless of whether patients had early or late stage tumors. One hypothesis is that in the case of low SES patients price of healthcare may prevent the use of high price therapies like chemotherapy and also insurance availability. The fact that the same trend in SES is also present when chemotherapy is recommended by NCCN strengthen this hypothesis. Differences in insurance coverage for treatment may explain differences in treatment response by SES. However, insurance coverage was not included within analyses as data on insurance provider represented less than 20% of patients within the CCR. One exception to this was observed in those patients that received no surgery and no radiation. In these patients low socio economic status was predictive of being less likely to have chemotherapy withheld. This flipping of the effect of the SES predictor could be attributed to access to insurance such as Medicaid which is only available to low SES patients. Currently the CCR has very limited data collected on the insurance status of oral cancer patients. Further research is needed exploring the relationship between SES and chemotherapy use.

This study could not identify a significant effect of racial group when predicting assignment of chemotherapy for early and late stage patients. Past attributions of difference in survival by race in oral cancer may be due in part to the distribution of socioeconomic status and access to care across racial groups. Lack of significance in racial group predictors indicate that disparities in overall survival between racial groups are not driven by disparities in chemotherapy assignment by race. It is possible that these differences in survival could be attributed to disproportionate distribution of

comorbidities by race that have been observed in past studies. Currently the CCR does not track comorbid status, access to this data would improve the ability of researchers to detect if comorbid status does contribute to racial disparities in survival.

Study Limitations

Multiple limitations should be addressed when assessing the results presented in this study. This study was unable to control for variables that were not reported by CCR, this could influence survival and compliance with NCCN recommended care. Comorbidities, tumor gene expression, chemotherapy dose, chemotherapy type, and dose intensity, Gy of radiation, and medication use were not included within CCR data. An analysis of the surgical techniques applied during treatment were not available at the time of this study. It is possible surgical approach and other aspects of technique could better define the effect of surgery on treatment response. The use of chemotherapy to promote quality of life was also not able to be measured as there are no measures within the CCR of which patients utilize chemotherapy to diminish suffering while receiving hospice care. The implications of insurance provider coverage was also not able to be assessed in this study. Although insurance provider data does exist for some patients within CCR this data is available in less than 20% of patients and is confounded by the period in which care was received as treatment coverage may vary from year to year for each insurer and between insurers. A stratified analysis taking insurance provider data into account would severely limit the power to detect any significant effects of treatment on survival. Lastly, this study did not consider effects of provider metrics (number of patients seen, number years of experience) and institution

type that have been the focus of other CCR studies examining the effect of NCCN compliance on outcomes in ovarian cancer[28]. Further studies of these variables may enhance interpretation of results reported in this study. These limitations do not detract from the reported results, and their ability to provide a unique contribution in validating NCCN recommendations in oral cancer both in those patients that receive recommendation for chemotherapy by NCCN and those that do not.

Conclusion

It has been shown that in those patients receiving care in compliance with NCCN guidelines, there is no significant difference in survival by race. It is clear that special attention should be paid to furthering the guidance provided by NCCN in applying recommended treatment equally across race, and thus minimizing differences in overall survival across racial groups. Chemotherapy treatment is not associated with improved survival of all oral cancer patients. The majority of patients (consisting of low stage node negative tumors) were found to have poorer survival when prescribed chemotherapy. Patients that do not receive surgery with early clinical stage node negative tumors were not shown to receive benefit when assigned to a chemotherapeutic regimen by their providers. This illustrates the idea that offering patients more treatment does not necessarily lead to better treatment outcomes. Greater guidance should be provided by NCCN in regards to which patients would receive the greatest benefit from chemotherapy, and also specifically for which patients' chemotherapy use has been associated with decreased survival.

Table 2.1. Study Population Characteristics					
Characteristic					
	All Patients	Surgery and Radiation *	Surgery*	Radiation *	No Surgery No Radiation*
Total	N=37,985	N=5,447	N=2126	N=10,178	N=18,289
Age at diagnosis					
Younger than 45	3834(10.0 %)	237(4.35 %)	113(5.31 %)	1160(11.3 %)	2225(12.1 %)
45-54	5860(15.4 %)	961(17.6 %)	270(12.6 %)	1755(17.2 %)	2627(14.3 %)
55-69	14036(36. 9%)	2435(44. 7%)	826(38.8 %)	3987(39.1 %)	6127(33.5 %)
70 or Older	13275(34. 9%)	1660(30. 4%)	864(40.6 %)	3022(29.6 %)	6851(37.4 %)
missing	980(2.57 %)	154(2.82 %)	53(2.49%)	254(2.49 %)	459(2.50 %)
Total	37985(100 %)	5447(100 %)	2126(100 %)	10178(10 0%)	18289(100 %)
Mean, Median	63.05, (13.72)	62.89,(11 .6)	62.5, (14.4)	62.2, (14.7)	63.6, (15.1)
Stage					
I	18062(47. 5%)	842(15.4 %)	437(20.5 %)	2916(28.6 %)	13418(73. 3%)
II	5406(14.2 %)	849(15.5 %)	344(16.1 %)	1735(17.0 %)	2136(11.6 %)
III	4476(11.7 %)	1181(21. 6%)	249(11.7 %)	2051(20.1 %)	826(4.51 %)
IV	8590(22.6 %)	2327(42. 7%)	842(39.6 %)	3319(32.6 %)	1557(8.51 %)
Unspecified	1451(3.81 %)	248(4.55 %)	254(11.9 %)	157(1.54 %)	352(1.92 %)
Total	37985(100 %)	5447(100 %)	2126(100 %)	10178(10 0%)	18289(100 %)
Grade					
1	9158(24.1 %)	659(12.0 %)	340(15.9 %)	1429(14.0 %)	6336(34.6 %)
2	18186(47. 8%)	2381(43. 7%)	995(46.8 %)	4785(47.0 %)	9088(49.6 %)
3	9311(24.5 %)	2267(41. 6%)	723(34.0 %)	3231(31.7 %)	2530(13.8 %)
4	1330(3.50 %)	140(2.57 %)	68(3.19%)	733(7.20 %)	335(1.83 %)

Total	37985(100%)	5447(100%)	2126(100%)	10178(100%)	18289(100%)
Tumor Thickness					
>25 cubic mm	17728(46.6%)	1331(24.4%)	462(21.7%)	4341(42.6%)	11255(61.5%)
<26 cubic mm	12228(32.1%)	2558(46.9%)	911(42.8%)	4538(44.5%)	3591(19.6%)
Unknown	8029(21.1%)	1558(28.6%)	753(35.4%)	1299(12.7%)	3443(18.8%)
Total	37985(100%)	5447(100%)	2126(100%)	10178(100%)	18289(100%)
QUINYOST Socioeconomic Status					
1	4159(10.9%)	612(11.2%)	311(14.6%)	1098(10.7%)	1784(9.75%)
2	5299(13.9%)	734(13.4%)	307(14.4%)	1382(13.5%)	2500(13.6%)
3	6034(15.8%)	819(15.0%)	316(14.8%)	1549(15.2%)	2968(16.2%)
4	6073(15.9%)	770(14.1%)	292(13.7%)	1607(15.7%)	3049(16.6%)
5	6180(16.2%)	813(14.9%)	315(14.8%)	1644(16.1%)	3090(16.8%)
missing	10240(26.9%)	1699(31.1%)	585(27.5%)	2898(28.4%)	4898(26.7%)
Total	37985(100%)	5447(100%)	2126(100%)	10178(100%)	18289(100%)
Gender					
Male	23852(62.7%)	3936(72.2%)	1372(64.5%)	6481(63.6%)	10854(59.3%)
Female	14133(37.2%)	1511(27.7%)	754(35.4%)	3697(36.3%)	7435(40.6%)
Total	37985(100%)	5447(100%)	2126(100%)	10178(100%)	18289(100%)
Racial Group					
White	29130(76.6%)	4294(78.8%)	1541(72.4%)	7540(74.0%)	14325(78.3%)
Black	1971(5.18%)	389(7.14%)	214(10.0%)	550(5.40%)	673(3.67%)
Hispanic	3954(10.4%)	491(9.01%)	213(10.0%)	1262(12.3%)	1806(9.87%)
Asian	2338(6.15%)	231(4.24%)	114(5.36%)	762(7.48%)	1135(6.20%)
Other†	592(1.55%)	42(0.77%)	44(2.06%)	64(0.62%)	350(1.91%)
Total	37985(100%)	5447(100%)	2126(100%)	10178(100%)	18289(100%)
Chemotherapy Assignment					

Assigned	6881(18.1%)	3000(55.0%)	658(30.9%)	2201(21.6%)	499(2.72%)
Not Asigned	31104(81.8%)	2477(45.4%)	1468(69.0%)	7977(78.3%)	17790(97.2%)
Total	37985(100%)	5477(100%)	2126(100%)	10178(100%)	18289(100%)
NCCN Chemotherapy Recommendation					
No Recommendation Given	23521(61.9%)	1716(31.5%)	783(36.8%)	4669(45.8%)	15561(85.0%)
Recommended	12048(31.7%)	2927(53.7%)	967(45.4%)	5210(51.1%)	2334(12.7%)
Missing	2416(6.36%)	804(14.7%)	376(17.6%)	299(2.93%)	394(2.15%)
Total	37985(100%)	5447(100%)	2126(100%)	10178(100%)	18289(100%)

SD, standard deviation; NOS, not otherwise specified. Data are n (%) unless otherwise specified. *1945 Patients missing data on surgical treatment. Of 1945 missing 543 are missing status on NCCN recommendation, 792 are not recommended to receive chemotherapy by NCCN and are missing surgery status, 610 are recommended to receive chemotherapy by NCCN and are missing surgery status.

Table 2.2 Multivariable Logistic Regression Analysis of Variables Associated With Withholding Chemotherapy in Patients with Node Positive Disease Stage III and Later

	All Patients(n=12048)			Surgery and Radiation (n=2927)			Radiation Alone (n=5210)			Surgery Alone (n=967)			No Radiation No Surgery (n=2334)		
Factors	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value
Grade Reference G4															
G3	1.25	1-1.4	<.0001	1.01	1.15-1.69	0.57	1.39	1.18-1.85	0.0005	0.74	.45-1.2	0.23	1.63	1.07-2.47	0.002

QUINYOST Socioeconomic Status															
1	1.54	1.21-1.95	0.0003	1.65	1.20-2.3	0.03	1.68	1.63-3.48	<.0001	1.7	.7-4.1	0.19	0.7	.35-1.4	0.37
2	1.38	1.1-1.7	0.0003	1.57	1.15-2.2	<.001	1.58	1.32-2.72	0.0005	0.63	.3-1.3	0.23	1.01	.41-2.02	0.96
3	1.24	1-1.5	0.03	1.59	.86-1.5	<.001	1.14	.95-1.85	0.63	1.4	.66-3.0	0.36	0.82	.4-1.6	0.58
4	1.26	1-1.5	0.02	1.74	.92-1.62	<.001	1.22	1.07-2.08	0.01	1.2	.57-2.4	0.65	0.6	.3-1.2	0.13
Age (compared to those 70 or older)															
Younger than 45	0.34	.37-.59	<.0001	0.34	.18-.65	0.001	0.39	.28-.56	<.0001	0.24	.07-.77	0.01	0.26	.13-.51	<.0001
45-54	0.34	.27-.41	<.0001	0.3	.20-.44	<.0001	0.38	.28-.51	<.0001	0.26	.12-.53	0.0003	0.34	.18-.61	0.0004
55-69	0.47	.4-.55	<.0001	0.38	.28-.52	<.0001	0.56	.43-.73	<.0001	0.45	.25-.77	0.004	0.48	.27-.84	0.01
Gender															
Male	0.7	.62-.85	<.0001	0.67	.50-.90	0.008	0.76	.61-.94	0.01	1.02	.61-1.7	0.91	0.61	.38-.96	0.03
Racial Group															
African American	1	.8-1.4	0.59	1.5	.93-2.4	0.18	1.19	.78-1.8	0.41	0.36	.15-.85	0.02	1	.42-2.6	0.89
Hispanic	1.4	1.2-1.6	0.03	0.91	.57-1.4	0.67	0.69	.51-.92	0.01	1.35	.58-3.08	0.48	0.42	.40-1.3	0.31
Asian	0.7	.5-.9	0.01	0.94	.50-1.7	0.85	0.72	.50-1.05	0.08	0.78	.23-2.6	0.69	0.37	.22-.82	0.01
Other	0.93	.3-2.1	0.84	1.17	.27-5.0	0.98	0.98	.27-3.4	0.97	0.4	.02-6.5	0.51	0.19	.05-4.5	0.54
Radiation Treatment															
No Radiation Treatment	2	1.7-2.4	<.0001	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Surgical Treatment															
No Surgical	6.6	5.7-7.6	<.0001	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Treatment															
Tumor Size (reference Tumor>25mm)															
Tumor is <26 mm	1.4	1.2-1.6	<.0001	1.35	1.02-1.90	0.03	1.34	1.1-1.7	0.004	1.3	.68-2.4	0.43	1.32	.9-2.2	0.12
*610 patients were missing information regarding surgical treatment. *The total number of late stage node positive patients added to the total number of early stage node positive patients does not sum to 37,985 due to the fact that 2,416 patients were missing information on nodal status or clinical stage.															

Table 2.3 Multivariable Logistic Regression Analysis of Variables Associated With Withholding Chemotherapy in Patients with Node Negative Disease Stage III and Earlier

	All Patients(n=23521)			Surgery and Radiation (n=1716)			Radiation Alone (n=4669)			Surgery Alone (n=783)			No Radiation No Surgery (n=15561)		
Factors	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value
Grade Reference G4															
G3	1.9	1.5-2.3	<.0001	1.7	1.2-2.3	0.001	1.9	1.4-2.5	<.0001	0.7	.35-1.4	0.32	2.2	1.4-3.7	0.0007
QUINYOST Socioeconomic Status															
1	0.94	.72-1.2	0.63	1.1	.7-1.9	0.62	1.03	.65-1.6	0.89	0.76	.27-2.1	0.6	0.58	.31-1.1	0.09
2	1.17	.91-1.5	0.24	1.5	.9-2.4	0.1	1.18	.77-1.8	0.43	0.95	.34-2.6	0.92	0.87	.5-1.7	0.68
3	1.2	.9-1.6	0.14	1.4	.9-2.3	0.12	1.26	.83-1.9	0.27	0.46	.19-1.1	0.09	1.6	.8-3.4	0.17
4	1.3	1-1.7	0.03	1.8	1.1-2.9	0.01	1.53	.99-2.3	0.05	0.4	.16-1.0	0.05	1.16	.6-2.2	0.64
Age (compared to those 70 or older)															
Younger than 45	0.54	.4-.8	0.0003	0.37	.17-.83	0.01	0.61	.38-.98	0.04	0.66	.16-2.6	0.56	0.44	.2-.9	0.03

45-54	0.48	.4-.6	<.0001	0.34	.20-.56	<.0001	0.58	.38-.89	0.01	0.57	.2-1.4	0.24	0.46	.2-.9	0.04
55-69	0.53	.43-.66	<.0001	0.46	.31-.66	<.0001	0.76	.54-1.08	0.13	0.59	.3-1.1	0.12	0.28	.16-.5	<.0001
Gender															
Male	0.8	.76-1.1	0.42	0.82	.58-1.16	0.27	0.85	.63-1.15	0.3	1.3	.73-2.5	0.31	1.1	.7-1.7	0.55
Racial Group															
African American	1	.69-1.4	0.97	1.8	.9-3.6	0.09	1.15	.61-2.17	0.65	0.65	.2-1.8	0.41	0.48	.2-.9	0.04
Hispanic	0.82	.6-1.1	0.19	0.65	.36-1.2	0.17	1.1	.68-1.7	0.69	0.37	.15-.92	0.03	0.8	.4-1.5	0.51
Asian	0.82	.6-1.2	0.3	0.82	.40-1.7	0.58	0.92	.55-1.55	0.76	0.52	.16-1.0	0.37	0.8	.3-1.9	0.64
Other	0.52	.2-1.3	0.18	<.001	<.001-->999	0.98	0.28	.06-1.3	0.11	>999	<.001-->999	0.98	1.1	.2-8.1	0.92
Radiation Treatment															
No Radiaiton Treatment	3.43	2.8-4.2	<.0001	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Surgical Treatment															
No Surgical Treatment	5.6	4.6-6.8	<.0001	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Tumor Size (reference Tumor>25mm)															
Tumor is <26 mm	3.4	2.9-4.1	<.0001	2.1	1.45-3.04	<.0001	1.7	1.3-2.3	0.0002	3	1.5-6.0	0.002	2.2	1.4-3.5	0.0007
792 Patients were missing information on surgical treatment. *The total number of late stage node positive patients added to the total number of early stage node positive patients does not sum to 37,985 due to the fact that 2,416 patients were missing information on nodal status or clinical stage.															

Table 2.4. Cox Proportional Hazards Model for Oral Cancer Patients with Node Positive Disease Stage III and Later															
	All Patients(n=12048)			Surgery and Radiation (n=2927)			Radiation Alone (n=5210)			Surgery Alone (n=967)			No Radiation No Surgery (n=2334)		
Factors	HR	95% CI	p-value	HR	95% CI	p-value	HR	95% CI	p-value	HR	95% CI	p-value	HR	95% CI	p-value
Clinical Stage (Reference Remote Disease)															
Regional by Lymph Nodes	0.67	.09-4.8	0.69	0.5	.38-.64	<.0001	0.26	.15-.46	<.0001	0.27	.2-.5	<.0001			
Regional by Direct Extension and Lymph Nodes	0.67	.62-.73	<.0001	0.6	.5-.7	<.0001	0.54	.39-.73	0.0001	0.54	.4-.7	0.0001	0.9	.7-1.00	0.07
Regional NOS	0.49	.81-.94	0.0002	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.1	.9-1.3	0.3

Grade Reference G4													0.6	.2-1.6	0.3
G3	1.03	.97-1.09	0.23	1.3	1.1-1.5	0.0005	0.97	.71-1.3	0.03	1	.7-1.3	0.85	0.8	.6-.9	0.0001
Tumor Size (reference >25mm)															
Tumor less than 26mm	0.7	.65-.74	<.0001	0.65	.5-.8	<.0001	0.77	.54-1.1	0.17	0.8	.5-1.1	0.17	0.7	.6-.8	<.0001
QUINYOST Socioeconomic Status															
1	1.4	1.2-1.5	<.0001	1.9	1.4-2.4	<.0001	1.5	.88-2.4	0.13	1.5	.9-2.4	0.13	1.2	.9-1.5	0.14
2	1.3	1.2-1.4	<.0001	1.7	1.3-2.1	<.0001	1.5	.95-2.3	0.07	1.5	.9-2.3	0.07	1.5	1.2-1.8	0.0003
3	1.3	1.1-1.4	<.0001	1.5	1.2-1.9	0.0007	1.4	.91-2.2	0.11	1.4	.9-2.2	0.11	1.2	.9-1.4	0.13
4	1.2	1.05-1.25	<.0001	1.5	1.2-1.8	0.001	1.2	.7-1.9	0.49	1.2	.7-1.9	0.46	1.1	.9-1.3	0.34
Gender															
Male	0.98	.92-1.05	0.71	0.95	.8-1.1	0.63	0.9	.7-1.3	0.82	1	.7-1.3	0.82	1	.9-1.2	0.64
Racial Group															
African American	1.3	1.1-1.5	<.0001	1.2	.9-1.5	0.21	1.2	.7-1.9	0.44	1.2	.7-1.9	0.44	1.3	.9-1.8	0.08
Hispanic	1	.9-1.1	0.95	1.1	.8-1.4	0.36	1.1	.6-1.9	0.28	1.2	.7-1.9	0.59	0.8	.6-1.0	0.03
Asian	0.91	.8-1.0	0.18	0.8	.6-1.2	0.33	1.1	.5-2.3	0.85	1	.5-2.3	0.85	0.8	.6-1.1	0.19
Other	0.92	.6-1.4	0.71	1.4	.6-3.2	0.43	0.6	.1-3.5	0.61	0.6	.1-3.5	0.61	0.5	.2-1.2	0.11
Chemotherapy Treatment Status															
Received Chemotherapy Treatment	0.98	.9-1.05	0.67	0.7	.6-.8	<.0001	0.9	.7-1.2	0.46	0.9	.7-1.2	0.46	1.2	.9-1.5	0.09

Radiation Treatment Status															
No Radiation Treatment	1.272	1.2-1.4	<.0001	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Surgical Treatment Status															
No Surgical Treamen t	0.88	.8-.96	0.003	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
610 patients are missing surgical treatment. *The total number of late stage node positive patients added to the total number of early stage node positive patients does not sum to 37,985 due to the fact that 2,416 patients were missing information on nodal status or clinical stage. ."NA" indicates that this strata is not applicable as tumors with the clinical stage in question do not exist within the treatment modality subgroup.															

Table 2.5. Cox Proportional Hazards Model for Oral Cancer–Specific Overall Survival in Patients with Node Negative Disease Stage III and Earlier

	All Patients(n=23521)			Surgery and Radiation (n=1716)			Radiation Alone (n=4669)			Surgery Alone (n=783)			No Radiation No Surgery (n=15561)		
Factors	HR	95% CI	p-value	HR	95% CI	p-value	HR	95% CI	p-value	HR	95% CI	p-value	HR	95% CI	p-value
Clinical Stage (Regional by Lymph Nodes (Node Negative))															
Localized	1.5	.9-2.6	0.1	4.2	1.3-14	0.01	1.3	.6-2.7	0.5	0.6	.5-.9	0.003	0.6	.2-1.6	0.3
Regional by direct Extension	2.1	1.3-3.5	0.003	4.6	1.4-15	0.01	1.6	.8-3.3	0.2	NA	NA	NA	0.9	.3-2.3	0.7
Tumor Grade (reference G3)															
<G3	0.8	.8-.9	<.0001	1.2	1-1.5	0.02	0.9	.8-1.1	0.6	0.9	.7-1.4	0.87	0.7	.6-.7	<.0001
Tumor Size (greater than 25cm)															
Less than 26 cm	0.7	.7-.8	<.0001	0.8	.7-1.0	0.02	0.8	.8-.9	0.0004	0.6	.4-.8	0.0003	0.7	.7-.8	<.0001
QUINYOST Socioeconomic Status															
1	1.4	1.3-1.5	<.0001	1.3	1-1.7	0.04	1.5	1.2-1.7	<.0001	1.6	.9-2.5	0.06	1.4	1.2-1.5	<.0001
2	1.2	1-1.2	<.0001	1.4	1-1.8	0.01	1.3	1.1-1.5	0.0008	1	.7-1.6	0.9	1.2	1.1-1.3	<.0001

3	1.1	1-1.2	0.0008	1.1	.9-1.4	0.5	1.1	1-1.3	0.17	1	.6-1.3	0.6	1.15	1.1-1.2	0.0003
4	1.1	1-1.1	0.04	1.5	1.2-1.9	0.001	1.2	1-1.3	0.02	0.9	.6-1.4	0.6	1	.9-1.1	0.68
Gender															
Male	1.1	1.1-1.2	<.0001	0.9	.7-1	0.18	1.2	1.1-1.3	0.0007	0.9	.7-1.2	0.5	1.2	1.2-1.3	<.0001
Racial Group															
African American	1.2	1-1.3	0.007	1.2	.9-1.7	0.2	1.1	.9-1.4	0.27	1.2	.7-2.1	0.4	1	.9-1.2	0.7
Hispanic	0.76	.7-.82	<.0001	0.7	.5-1	0.08	0.7	.6-.8	<.0001	1	.6-1.7	0.9	0.8	.7-.9	<.0001
Asian	0.8	.7-.9	<.0001	1.1	.8-1.7	0.5	0.8	.7-1	0.05	1.3	.6-3.1	0.5	0.8	.7-.85	0.0002
Other	0.7	.6-.9	0.004	1.1	.2-4.9	0.9	1.1	.5-2.4	0.9	1.2	.3-4.2	0.8	0.7	.5-.9	0.003
Radiation Treatment Status															
No Radiation Treatment	0.9	.85-.95	0.0003	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Surgical Treatment Status															
No Surgical Treamen t	0.64	.59-.69	<.0001	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Chemotherapy Treatment Status															
Received Chemotherapy Treatment	1.5	1.4-1.7	<.0001	1	.8-1.2	0.9	1.4	1.2-1.7	<.0001	1.3	.9-2.0	0.16	1.9	1.5-2.3	<.0001
792 Patients were missing information on surgical treatment. *The total number of late stage node positive patients added to the total number of early stage node positive patients does not sum to 37,985 due to the fact that 2,416 patients were missing information on nodal status or clinical stage.															

Table 2.6 Clinical Stage and Nodal Status by Chemotherapy Assignment				
NCCN Chemotherapy recommendation	Measure	Chemotherapy		
		Received	Not Received	Total
Early Stage (Stage I-III Node Negative) Chemotherapy not recommended by NCCN	Frequency	1618	21903	23521
	Percent	4.55	61.58	66.13
	Row Pct	6.88	93.12	
	Col Pct	26.45	74.37	
Late Stage (Stage III-IV Node Positive) Chemotherapy Recommended by NCCN	Frequency	4499	7549	12048
	Percent	12.65	21.22	33.87
	Row Pct	37.34	62.66	
	Col Pct	73.55	25.63	
Total		6117	29452	35569
		17.2	82.8	100

*The total number of patients recommended to receive chemotherapy treatment by NCCN added to the total number of patients not recommended to receive chemotherapy treatment by NCCN does not sum to 37,985 due to the fact that 2,416 patients were missing information on NCCN recommendation. This missingness is attributable to either missing information on stage or nodal status which are used to code the NCCN recommendation variable.

Table 2.7. Patient Use of Chemotherapy by Anatomic Site, Stage and Nodal Status (n,% Patients Receiving Treatment Specific to Anatomic Site and Stage/Nodal Status)						
		Lip	Tongue	Floor of Mouth	Gum	Other Oral Cavity
Late Stage Node Positive	Chemotherapy received	33 (21.7 %)	2961 (49.1 %)	301 (26.6%)	597 (25.8 %)	178 (39.2%)
	Chemotherapy not received	119 (78.2 9%)	3081 (50.9 %)	827 (73.3%)	1715 (74.2 %%)	276 (60.8%)
Early Stage Node Positive	Chemotherapy received	33 (.7%)	852 (11.3 %)	160 (6.2%)	328 (6.3%)	92 (23.2%)
	Chemotherapy not received	4398 (99.3 %)	6692 (88.7 %)	2418 (93.8%)	4900 (93.7 %)	305 (76.8%)

Early stage refers to those patients that are Node negative and at stages I-III, Late stage refers to those patients that are node positive and at stages III-IV. Node positive refers to invasion of >1 lymph node.

Figure 2.1. Survival by Chemotherapy Treatment in Those Patients Recommended to Receive Chemotherapy by NCCN

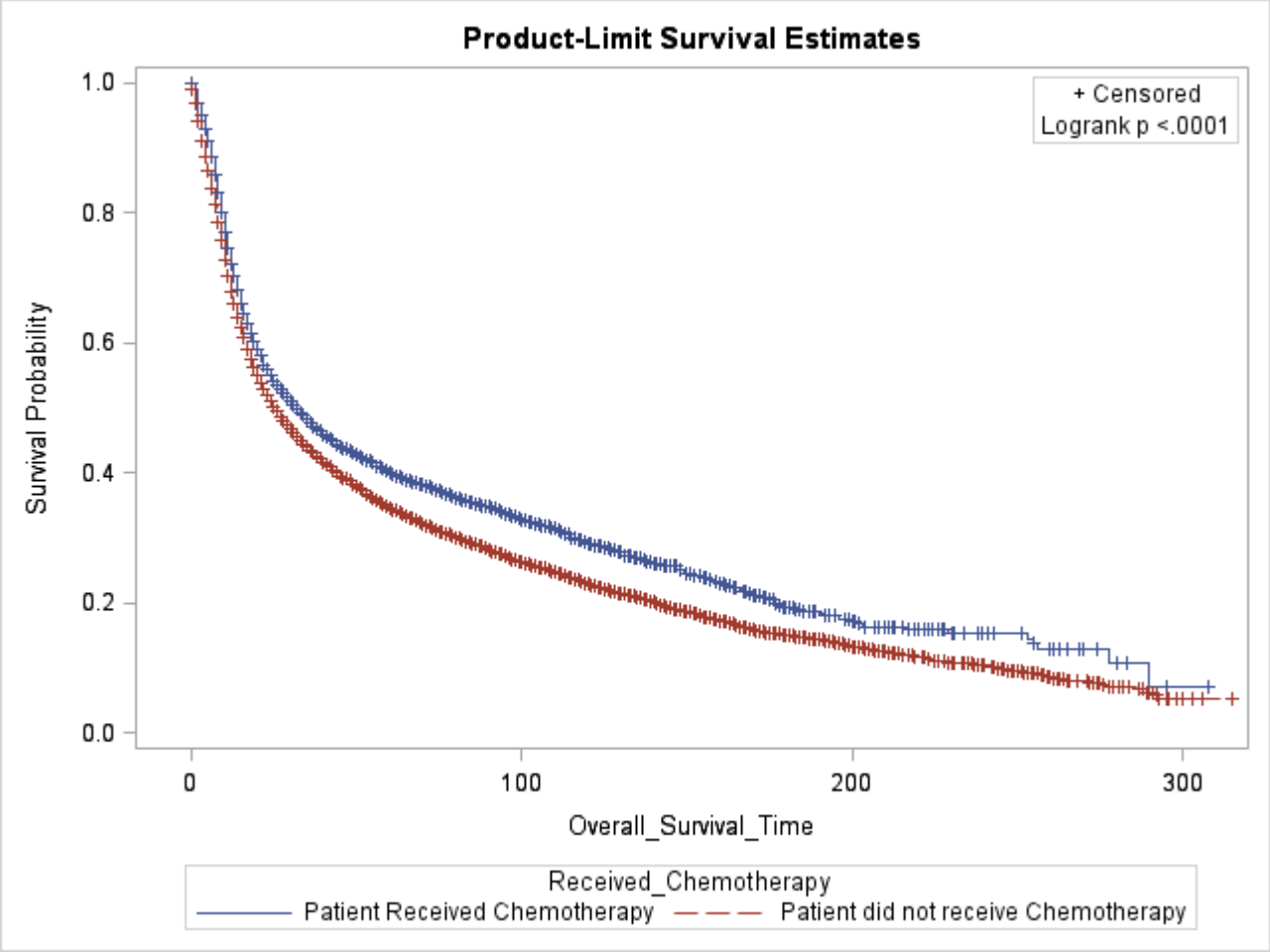


Figure 2.2. Survival by Chemotherapy Treatment in Those Patients Not Recommended to Receive Chemotherapy by NCCN

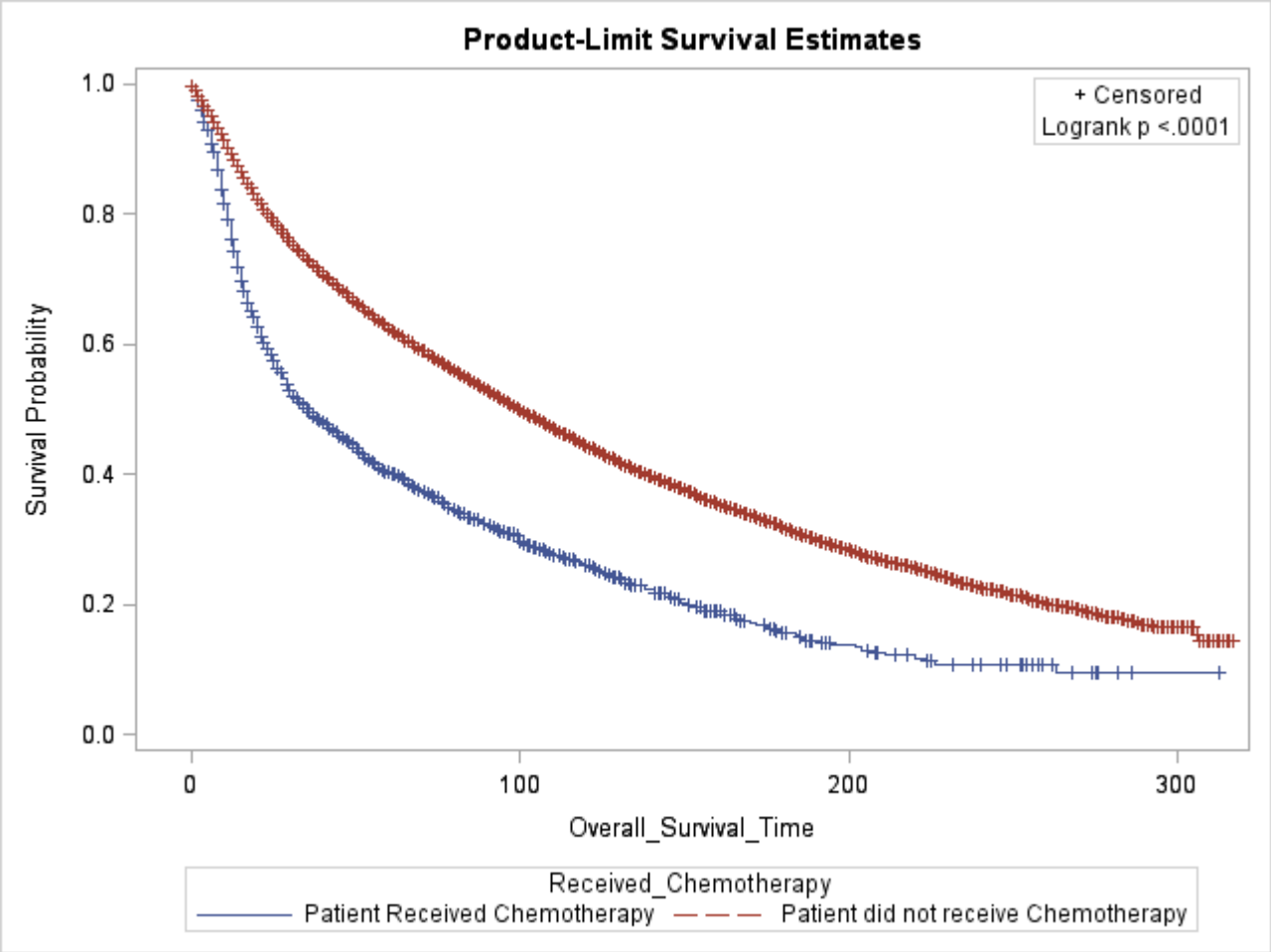


Figure 2.3. Study Exclusion Criteria Diagram

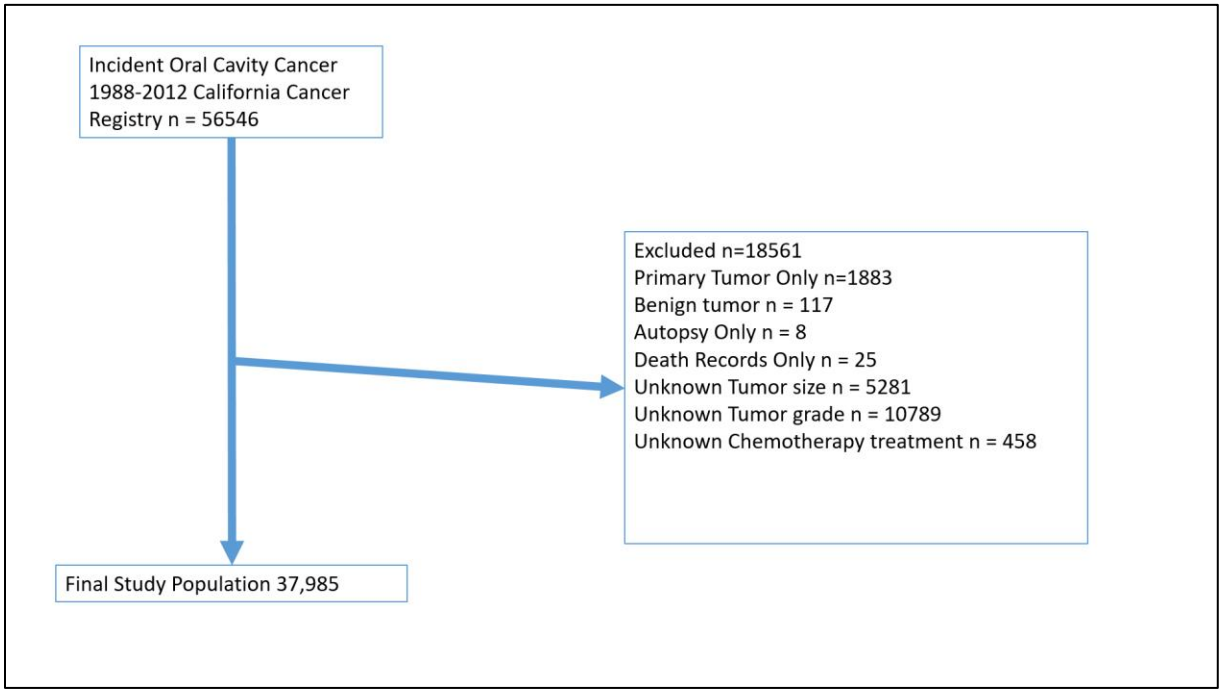
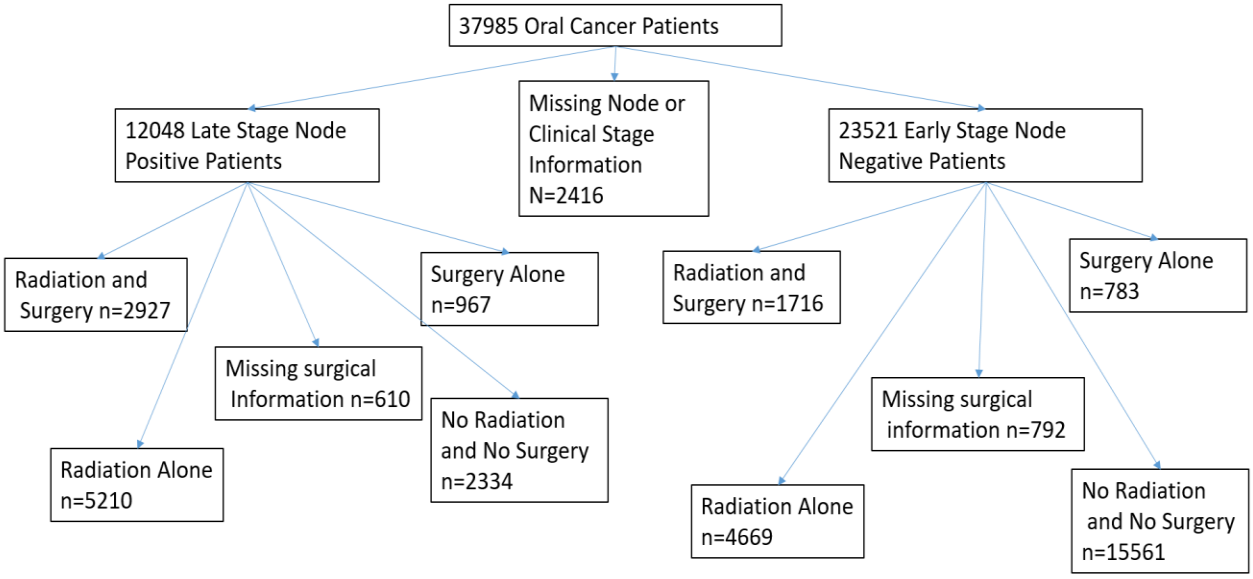


Figure 2.4. Role of NCCN Recommendation in Oral Cancer Study Population Subsets Diagram



Chapter 3

Identification of a Gene Expression Signature Predicting Survival in Oral Cavity Squamous Cell Carcinoma Using Monte Carlo Cross Validation

Introduction

Head and neck cancers are cancers of the upper airway and/or digestive tract found in the oral cavity, laryngeal, pharyngeal, oropharyngeal, and hypo-pharyngeal tissues. Head and neck cancers make up 3% of cancers diagnosed each year[1,2]. Head and neck cancer incidence has declined from 25 cases per 100,000 at risk in the 1990s to 15 cases per 100,000 at risk in the present day[3]. While the decrease in head and neck cancer incidence may be due to a drop in tobacco use [4,5], the mortality associated with these cancers has not changed significantly in the last twenty years[62]. Human Papilloma Virus (HPV) positive patients have been observed to have an improved survival and response to treatment when compared to HPV negative patients. However, these patients make up the minority of oral squamous cell cancers (OSCC) [7]. Thus, the decline in mortality could be attributed to decrease in smoking, increases in HPV positive cases, or other unknown mechanisms.

Few studies have identified a group of genes predicting treatment response in HPV-negative OSCC patients. To date, the most widely used molecular signature guiding head and neck squamous cell carcinoma (HNSCC) treatment is HPV status. HPV status can be measured directly through polymerase chain reaction analysis, or indirectly through cyclin-dependent kinase inhibitor 2A (CDKN2A) expression.

However, HPV preferentially infects oropharyngeal tissues which make up only 15% of HNSCC[30]. There have been multiple studies that have identified the genetic markers that improve prediction of overall survival when HPV status is known[35–38]. Unfortunately, there has been less focus on HPV-negative OSCC patients, an HNSCC subgroup that is known to respond significantly worse to treatment than patients with Oropharyngeal Squamous Cell Carcinoma (OPSCC)[35,38]. OSCC patients have been shown to be less likely to be HPV positive than Oropharyngeal cancer patients and thus are more reflective of the outcomes of HPV negative patients.

Past studies examining molecular signatures in OSCC have found that pathways in cell migration, cell-to-cell signaling and interaction, and cellular growth and proliferation are predictive of overall survival[63,64]. The keratin pathway is also notable in that it has been identified by several studies for its role in predicting the conversion of leukoplakia to malignant tumor, tumor progression, nodal stage, and overall response to treatment[65]. Of the OSCC studies listed the largest sample was 130 patients[64]. A common theme among the reported studies is low reproducibility in the genes identified as predictive of advanced disease or survival.

There has been much success in the production of site specific predictive models that draw upon the rich resource of data in the TCGA[66]. Models predicting survival in glioblastoma, colorectal, ovarian, and even head and neck cancer have drawn upon TCGA data in the past [67–71]. The 2015 study examining head and neck cancer data in the TCGA focused on gene mutations that were observed across all head and neck cancer patients and in those patients that tested HPV positive. While this study did

describe treatment response, it did not utilize gene expression data when conducting survival analyses. This study does draw upon gene expression data in the TCGA to produce an aggregated model that predicts survival across strata of tumor behavior, treatment regimen, and gender.

There are a host of methods that can be applied in the identification of a predictive molecular signature. When composing a signature that is predictive and prognostic, there are several quality checkmarks that must be addressed. Model building of any kind must go through an internal validation process where data is divided between test and training data. While model simplicity or complexity improve model usability, they are superseded in importance by measures of model performance[72]. Internal validation is an acceptable form of validation only when the test data set is completely untouched and no aspect of test data plays a part in model development. A drawback to splitting data in this way is the decrease in model efficiency due to the use of only a subset of the total data. One method addressing this inefficiency is to split a dataset into training and test data many times in a Monte Carlo validation (MCCV) or leave-one-out cross validation. These methods lead to nearly unbiased estimates of model performance (in the case of leave-one-out cross validation), and do not require sacrifice of sample size[73,74]. These methods have been applied by other studies in the successful identification of predictive models in many different types of cancer using leave-one-out cross validation [75–79] and MCCV [80,81]. The application of MCCV involves random sampling without replacement which means that subsets of the population with gene expression values with strong effect have a greater opportunity to

have that effect detected. MCCV differs from k-folds cross validation in that in MCCV an observation may be chosen to be included in a test set multiple times over the total number of iterations over all analyses opposed to one time in K-fold validation. MCCV is also viewed as a more conservative approach to cross validation as it overestimates the model prediction error in comparison to a k-fold cross validation which tends to underestimate prediction error[82]. External validation is an important and often costly task required for measuring a model's exportability. It is for this reason that robust internal validation measures should be adopted by those studies that lack the funding to carry out external validation in early stages of analysis.

Methods

Datasets

The Cancer Genome Atlas (TCGA) is a large, multi-dimensional, multi-center project compiling genomics data for over 29 cancer types into one central database[83]. TCGA contains clinical and demographic variables, gene expression profiling data, SNPs, protein expression, and methylation data. Clinical data on radiation dose, demographic variables, exposures (tobacco, alcohol, and HPV), chemotherapy type, and measures of overall and disease progression-free survival are included in the TCGA database (Table 1). Data accessed for this study were publicly available through the TCGA genomic data commons data portal. 523 head and neck cancer cases were downloaded from the data portal with all corresponding gene expression counts and corresponding clinical data. Of these 523 patients 313 OSCC patients were selected. 264 of the remaining 313 OSCC patients were included as only these patients possessed full survival data.

Differential Expression Analyses

Differential Gene Expression (DGE) analysis is a method of identifying genes that are expressed differently across time, tissue, and conditions, such as disease states[84]. This method of analysis uses fold change and significance criterion to select the genes in a molecular signature for predicting tumor phenotype, clinical subtype, or treatment response. All patients with cancer in tongue, lip, alveolar ridge, hard palate, floor of mouth, maxilla, and buccal mucosa were included. OSCC patients were the largest grouping of head and neck cancer patients and thus provided the most power to detect influential genetic pathways predicting treatment response. DGE analysis yielded a list of genes that are expressed differently between two strata. The strata used in this study were vital status within five years of follow-up. The TCGA RNA sequencing data were preprocessed with RSEM software, yielding normalized counts per million (CPM) gene expression counts[66]. The data were filtered to CPM >= 2, absolute fold change >=1.5, Fisher's exact p-value <= 0.05 and a false discovery rate <= 0.05. The list of genes produced by these filters was used to create a predictive signature comprised of 20 genes selected by the highest absolute log fold change value.

100 Runs of Differential Gene Expression Analysis using Monte Carlo Validation

A defining feature of MCCV is the random selection of observations into test and training sets across multiple iterations[85]. This study did require that some randomness be sacrificed, as a constant proportion of living and deceased were included at each iteration (opposed to a random proportion) to ensure that Cox regression survival analyses could be conducted. DGE analysis was repeated 100 times with a randomly selected (without replacement) set of 100 patients from the 264 total patients. Of the 100 patients selected in each iteration, 66 survived past 5 years and 34 were deceased prior to 5 years. At each iteration the top 20 genes with highest absolute fold change value were chosen to be placed in an additive Cox regression model predicting overall survival in OSCC patients. An AUC was produced for each of the signatures (comprised of 20 genes) created at each of the 100 iterations using the remaining 164 patients as a test set. The selected genes were aggregated to yield a table counting the number of times each gene met filter criteria over all the 100 iterations (Table 3.4). 100 iterations is double the number of iterations used in previous studies applying MCCV for similar purposes[86,87]. The number of genes within the final model was set at 40 to produce more robust estimates of survival than those models with 20 genes. The number of genes included in the signature did not exceed 40 as the model would not converge properly due to sample size restrictions. This application of MCCV has been used in the past to identify genetic predictors of disease status in breast cancer and Parkinson's disease[86,87]. This study applies a similar

method to identify those gene expression patterns that exert the greatest influence in predicting treatment response in OSCC.

The final aggregated model was comprised of counts per million for each gene in the final aggregated signature multiplied by a model weight. Once all 40 of the weighted CPM were summed across all 40 genes a risk score would be yielded indicating whether a patient would be set into high (>1.5) of low (<=1.5). The cutoff for high risk and low risk was set as the minimum difference between sensitivity and specificity on the ROC curve. This minimum value was identified using the pROC package in R[88]. In order to provide additional assurance that these results were not reached by chance alone, the study repeated the 100 signature validations using genes that were randomly selected from the 20530 genes in the dataset. The distribution of AUC across 100 runs of signatures based upon DGE analysis results were compared to the distribution of AUC derived from signatures comprised of genes that were randomly selected. To visualize these results, histograms were created by binning AUC by frequency (Figure 3.4).

Sensitivity of the Aggregated Signature

The sensitivity of the aggregated signature was validated by applying it to clinical subsets of all 264 test patients. Kaplan-Meier survival curves were used for this series of validation. Cox regression was used to determine the sensitivity of the aggregated signature when other variables were in the model. The cox model included race, gender, chemotherapy treatment, and tumor grade. Alcohol consumption and radiation variables were run in the model with dummy variables to address the effect of large

amount of missingness within these variables (145 missing variables in alcohol consumption, 45 missing variables for radiation dose). These variables were not found to have a significant impact on the estimates produced for the high risk scores and were excluded from the final cox regression model. Tumor Necrosis was excluded from the model due to the high amount of correlation with the gender variable which led to unstable estimates (There were no female patients with tumor necrosis < 15% present in the sample). Clinical stage was not included within this analysis due to the improved fit offered by the tumor grade variable, and both clinical stage and tumor grade were found to be nonsignificant when included within the model. Similar results for both tumor stage and clinical stage are not unexpected as tumor grade is a component of the clinical staging criteria. All analyses used age as a strata to prevent bias created by any skew in the distribution of age within each variable. Univariate cox regression was performed to provide context for multivariable analyses (Table 3.2).

Pathway Enrichment Analysis Methods

The R packages edgeR, and PA Reactome were used to conduct DGE and pathway enrichment analyses, respectively [89,90]. Pathway analysis tools and annotation databases were used to examine which pathways were enriched with the most frequently identified genes in the signatures produced over one hundred rounds of DGE. It is important to note that false discovery rate (FDR) produced by PA Reactome was not weighted for the frequency we observed genes to be significant over the 100 run DGE analysis, and thus the .05 FDR should be considered a conservative

threshold. A table of those pathways meeting a Fisher exact p-value threshold of .05 was included in the results (Table 3.3).

Results:

Differential Gene Expression

Each run of the DGE analysis identified differentially expressed genes based upon the gene expression values of randomly selected patients. An AUC reflecting the accuracy of each signature (each comprised of 20 genes) was recorded over 100 runs. These AUCs had a median of .84, max of .96, minimum of .65, mean of .83, and a standard deviation of .04. Similar analyses were performed on gene signatures of genes randomly selected from the 20530 genes in the dataset. The distribution of AUC for signatures made of randomly selected genes were median of .5, max of .63, minimum of .36, a mean of .5 and a standard deviation of .05 (Figure 3.4)

Differential gene expression analysis results were aggregated into a list of 40 of the most frequently identified differentially expressed genes included over all 100 runs of MCCV. (Table 3.4). When this molecular signature was tested in the dataset containing all patient data (n=264), it was found to correctly classify patient survival status, and it was found to have a specificity of 72%, sensitivity of 72%, and an area under the ROC curve of 75% (Figure3.1a, 3.1b). The distribution of patient demographics across risk scores can be viewed in (Table 3.1).

Validation of Aggregate Signature across Clinical Strata

This model was applied to subsets of the 264 patient test dataset. When overall survival difference was measured using all patients in the test set, it was found that there was a significant difference in patient survival outcomes when stratifying by the molecular signature risk score (p-value =2.6e-08) (Figure 3.1c). When stratifying by tumor grade, the signature was predictive of survival in those patients with high grade (Greater than G2) tumors and low grade (Less than G3) tumors (p-value .0008, 8.8e-06), respectively (Figure 3.2a, 3.2b).

The log rank survival by molecular signature risk score in only those patients receiving chemotherapy was (p-value=.002). The significance of difference by risk score in those patients not receiving chemotherapy was (p-value=.002) (Figure 3.3a, 3.3b). This signature continues to be predictive when all women were removed from the sample and the prediction of survival in men alone was tested (p-value=9.7e-07). However, this signature was not predictive in women and was found to be only marginally significant (p-value = 0.04) (Figure 3.3c, 3.3d).

Univariate and Multivariable Cox Regression

After adjusting for confounding variables, the signature risk score continued to be predictive of treatment response in both multivariable and univariate analyses (Table 3.2). High risk score was associated with an HR of 3.2 (95% CI 1.3 to 6.3, p-value<.0001) times greater odds of death when compared to patients with low risk score

in a univariate model. No significant effect was discovered when this model was applied to women alone. It was observed that both the signature and tumor necrosis lost effect size when performing multivariable adjustment. As this seemed indicative of possible correlation between the two variables, a Spearman correlation test was applied and yielded a 21% correlation significant with a (p-value = 9e-04). Our results showed that in addition to the signature risk score, gender and chemotherapy treatment were also predictive of overall survival.

Pathway Analysis Results

Significant pathways enriched with genes in the original signature were Interleukin, Calcitonin, ligand-gated ion channel transport, keratinization, and cornified envelope pathways (Table 3.3). There were no pathways that were enriched with greater than 2 genes from our signature. The most significantly enriched pathway was the ion channel transport pathway In total 11 genes from the 40 genes within the aggregated signature were identified as being enriched in the aforementioned pathways. The ligand gated ion channel transport pathway passed both fisher exact test and false discovery rate thresholds for significant enrichment (fisher's exact p-value =2.3 2-06, False discovery p-value=3.4 e-04). Genes within the ligand gated ion channel transport pathway were GLRA4 and HTR3C which were identified as significantly differentially expressed in 17% and 13% of the MCCV respective replications. All pathways listed in Table 3.3 meet a Fisher's exact p-value of .05 or less.

Discussion:

Interpretation of Signature Validation

The aggregated signature was shown to be predictive of treatment response in OSCC patients regardless of chemotherapy treatment status, or tumor grade. In addition to the identification of a signature that predicts overall survival in OSCC patients, this study also validated the use of Monte Carlo cross validation in producing gene signatures that are more likely to be reproduced across multiple studies. This method can be adopted by other researchers that wish to apply free and publicly available data to the testing of hypotheses in a manner that has the greatest likelihood of reproducibility across datasets.

Interpretation of Pathway Enrichment

The ion gate channel pathway was one of the pathways enriched with genes in the aggregated signature identified in this study. Ion gate channel pathway genes within the aggregated signature that were found to be significantly enriched were 5-Hydroxytryptamine Receptor (serotonin receptor) (HTR3C) and Glycine Receptor Alpha (4GLRA4). HTR3C has also been reported to be associated with other upper GI cancers such as esophageal adenocarcinoma [91]. Other Ion channel regulators like voltage-gated potassium channel Kv3.4 mRNA expression have been found to affect the progression of OSCCs, and inhibition of Kv3.4 inhibits growth of OSCC [92–94]. POU5F1, OCT4, SOX2, NANOG gene repression pathways were also found to be significantly enriched. These genes play a role in chemosensitivity to platinum based chemotherapies[95,96]. The keratin pathway is also notable in that it has been

identified for its role in predicting the conversion of leukoplakia to malignant tumor[97,98]. The MCCV approach did not detect all pathways typically associated with the development of OSCC. Pathways associated with HPV negative OSCC development include AKT, JNK, IL-6/STAT3, ILK, RAS, MAPK/ERK, p38/PAK, TGFβ, PI3K/mTOR and WNT signaling. The research questions focused upon by this study were which pathways were associated with treatment response thus, pathways associated with progression were not identified. Evidence of supporting literature is provided (Table 3.4) in a matrix of gene names and search terms related to OSCC, head and neck cancer, and cancer treatment response produced by Pubmatrix[99] (Table 3.4). The Pubmatrix results show that 65% of genes identified in this study are supported by existing literature reporting these genes roles in treatment response, survival, and progression.

Strengths and Limitations

This study had several limitations, TCGA data are known to be biased towards patients with later stage cancers with tumor sizes that are greater than 200g[71,100]. Additionally, samples in TCGA are contributed by multiple academic medical centers where collection methods may vary. When studying rare cancers it is common to have analysis curtailed by sample size, which is the limitation that this study hopes to specifically address through the application of MCCV. OSCC occurs more often in men than women and thus women only make up approximately 1/3 of our sample. The Monte Carlo validation approach is well suited to address these sample size limitations and is meant to serve as a model for other studies utilizing similar datasets. A

drawback of the MCCV approach is that it necessitates discarding signatures identified as predictive in single runs. Such sacrificed signatures may indeed point to true biological mechanisms which the other iterations of analysis did not detect due to their unique mix of patients. MCCV is designed to exclude all but the strongest effects. In many cases a combination of weak effects of genes may produce a predictive signature that can classify with accuracy but makes interpretation of biological mechanisms difficult. This study provides support for greater adoption of MCCV when conducting genomic or transcriptomic research in less common cancers.

Conclusion

The role of ion gate channel pathway in OSCC and its role in a molecular signature predicting treatment response is supported by this study. The ion channel gate pathway was the only pathway to pass both fisher exact test and false discovery rate significance thresholds. These results provide evidence that applying a MCCV approach to DGE model creation is a suitable method to control variability in results when using heterogeneous datasets, and offers a method of validation prior to devoting time and funding required for additional sequencing. The robustness of this signature was supported by the finding that the distribution of AUC for random signatures and signatures selected through MCCV were completely separate. Those researchers adopting heterogeneous datasets combined over multiple studies must address issues of result variability if they truly wish to contribute to the advancement of this field. This study describes and validates one approach that may be applied towards this goal.

Table 3.1 Patient Demographics Stratified by Molecular Signature

Characteristics	ALL (264, 100%)	Low Risk (n=151, 57%)	High Risk (n=113, 42%)
Vital Status			
Alive	189	(130, 86.6%)	(59, 52.2%)
Deceased	75	(21, 13.9%)	(54, 47.7%)
Age			
Age greater than 60	152	(85, 56.2%)	(67, 59.2%)
Age less than 61	112	(66, 43.7%)	(46, 40.7%)
Gender			
Female	88	(55, 36.4%)	(33, 29.2%)
Male	176	(96, 63.5%)	(80, 70.7%)
Tumor Grade			
G1	34	(19, 12.6%)	(15, 13.3%)
G2	153	(92, 61.3%)	(61, 54.4%)
G3	59	(29, 19.3%)	(30, 26.7%)
G4	5	(3, 2.0%)	(2, 1.7%)
GX	11	(7, 4.6%)	(4, 3.5%)
Race			
White	224	(127, 86.3%)	(97, 88.1%)
Not White	33	(20, 13.6%)	(13, 11.8%)
Clinical Stage			
Stage I	8	(4, 2.7%)	(4, 3.6%)
Stage II	57	(28, 19.1%)	(29, 26.1%)
Stage III	58	(38, 26.0%)	(20, 18.0%)
Stage IVA	126	(71, 48.6%)	(55, 49.5%)
Stage IVB	6	(4, 2.7%)	(2, 1.8%)
Stage IVC	2	(1, .6%)	(1, .9%)
Alcoholic Drinks>2 consumed per day			
TRUE	57	(37, 50.6%)	(20, 41.6%)
FALSE	64	(36, 49.3%)	(28, 58.3%)
History of Smoking			
TRUE	195	(113, 74.8%)	(82, 72.5%)
FALSE	69	(38, 25.1%)	(31, 27.4%)
Tumor Necrosis Greater than or equal to 15%			
TRUE	115	(60, 41.0%)	(55, 50.4%)
FALSE	140	(86, 58.9%)	(54, 49.5%)
Radiation >66 Gy			

TRUE	23	(14, 11.4%)	(9, 9.2%)
FALSE	196	(108, 88.5%)	(88, 90.7%)
Receiving Chemotherapy			
TRUE	95	(60, 39.7%)	(35, 30.9%)
FALSE	169	(91, 60.2%)	(78, 69.1%)

Table 3.1."Chemotherapy**"** is not specific to a given chemotherapeutic agent. This merely reflects whether a patient was assigned to chemotherapy treatment or not. History of Smoking stratifies patients into "never" or "ever" smokers. High Grade includes G1 and G2 patients, while low grade includes G3, G4, GX tumor grades. Not all characteristics total to 264 as some variables were incomplete (Tumor Grade NA=2, Clinical Stage NA=7, alcohol consumption per day NA=143, Tumor Necrosis NA=9, Radiation NA=45)

Table 3.2 Univariate and Multivariable Cox Regression Analyses

	Univariate			Multivariable		
Characteristic	HR	95% Confidence Interval	p-value	HR	95% Confidence Interval	p-value
No Smoking History	0.7	0.4-1.2	0.24	0.6	0.1-3.2	0.5
Female Gender	0.4	0.2-0.7	0.002	0.4	0.2-.07	0.004
Tumor Grade <2	0.7	.5-1.2	0.23	0.7	0.4-1.2	0.6
Caucasian Race	1.0	0.4-1.9	0.96	1.0	0.5-2.16	0.90
Chemotherapy Not Received	1.9	1.1-3.4	0.01	1.9	1.1-3.5	0.01
High Risk Signature	3.3	1.9-5.5	<.0001	3.25	1.3-6.3	<.0001

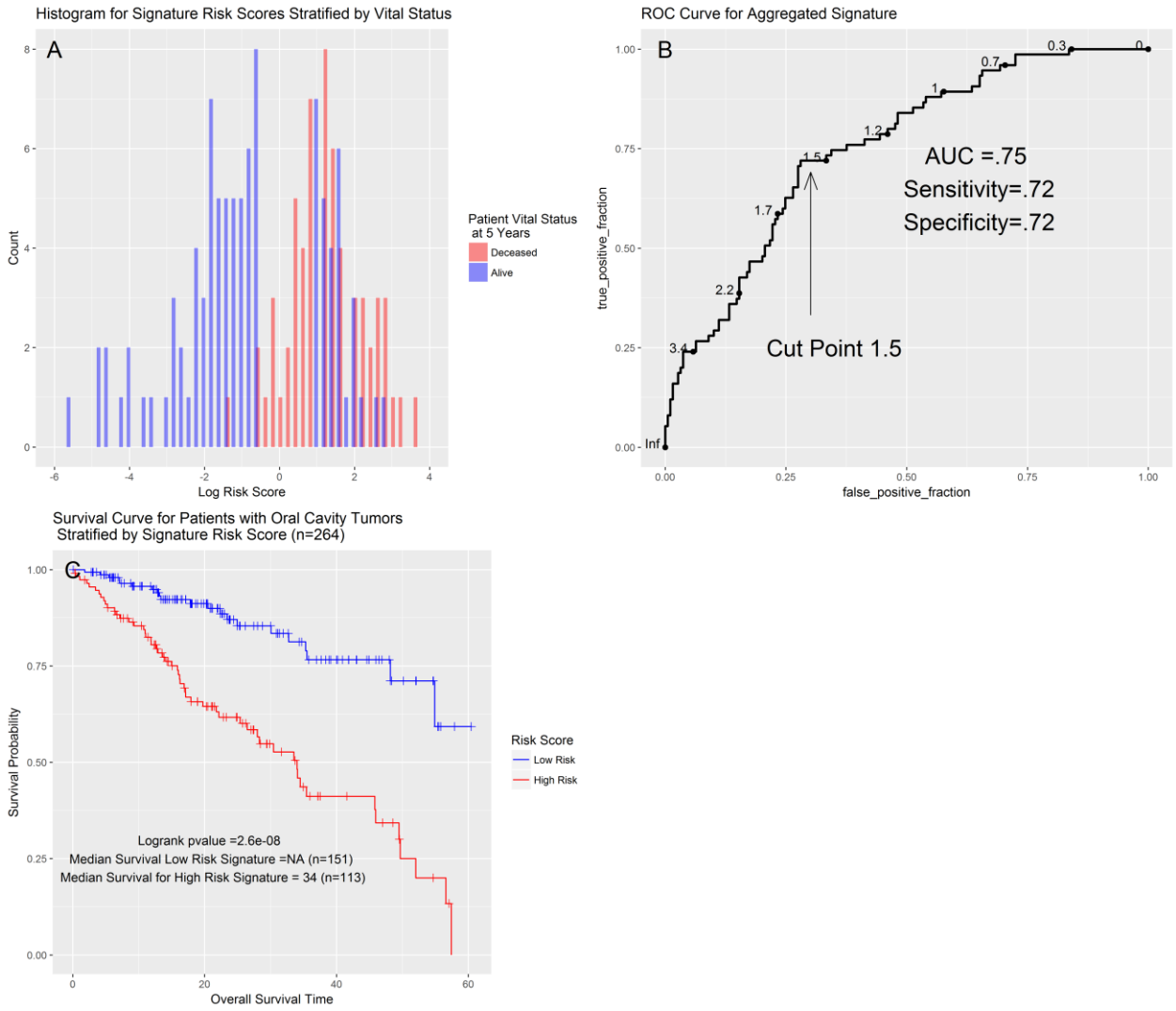
Table 3.2. Univariate and Multivariable Cox Regression adjusting for pertinent clinical strata. All p-values less than .05 are considered significant. Radiation and Alcohol not included in analyses within table due to high number of missing observations. Tumor Necrosis removed from table due to the fact that there were 0 female patients with tumor necrosis > 15%. Clinical Stage not included as Tumor Grade and stage correlated at 39.5%. All Analyses were age stratified.

Table 3.3. Pathway Analysis of Aggregated Signature

Pathway Name	Number of Genes from Aggregate Signature in Pathway	Total Number of Genes in Pathway	Fisher's Exact p-value	Aggregated Signature Genes Found in Pathway
Ligand-gated ion channel transport	2	33	2.27E-06	HTR3C;GLRA4
Defective pro-SFTPC causes pulmonary surfactant metabolism dysfunction 2 (SMDP2) and respiratory distress syndrome (RDS)	1	2	0.005	SFTPC
Assembly of active LPL and LIPC lipase complexes	1	30	0.01	FGF21
Surfactant metabolism	1	52	0.01	SFTPC
Formation of the cornified envelope	2	130	0.01	KRT38;KRT72
Defective ABCA3 causes pulmonary surfactant metabolism dysfunction type 3 (SMDP3)	1	9	0.02	SFTPC
Regulation of signaling by NODAL	1	12	0.03	LEFTY2
Calcitonin-like ligand receptors	1	11	0.03	CALCR
Plasma lipoprotein remodeling	1	54	0.03	FGF21
Class B/2 (Secretin family receptors)	2	99	0.04	CALCR;GLP2R
Keratinization	2	218	0.04	KRT38;KRT72
POU5F1 (OCT4), SOX2, NANOG repress genes related to differentiation	1	10	0.04	CDX2
Interleukin-4 and 13 signaling	1	212	0.04	IL17A

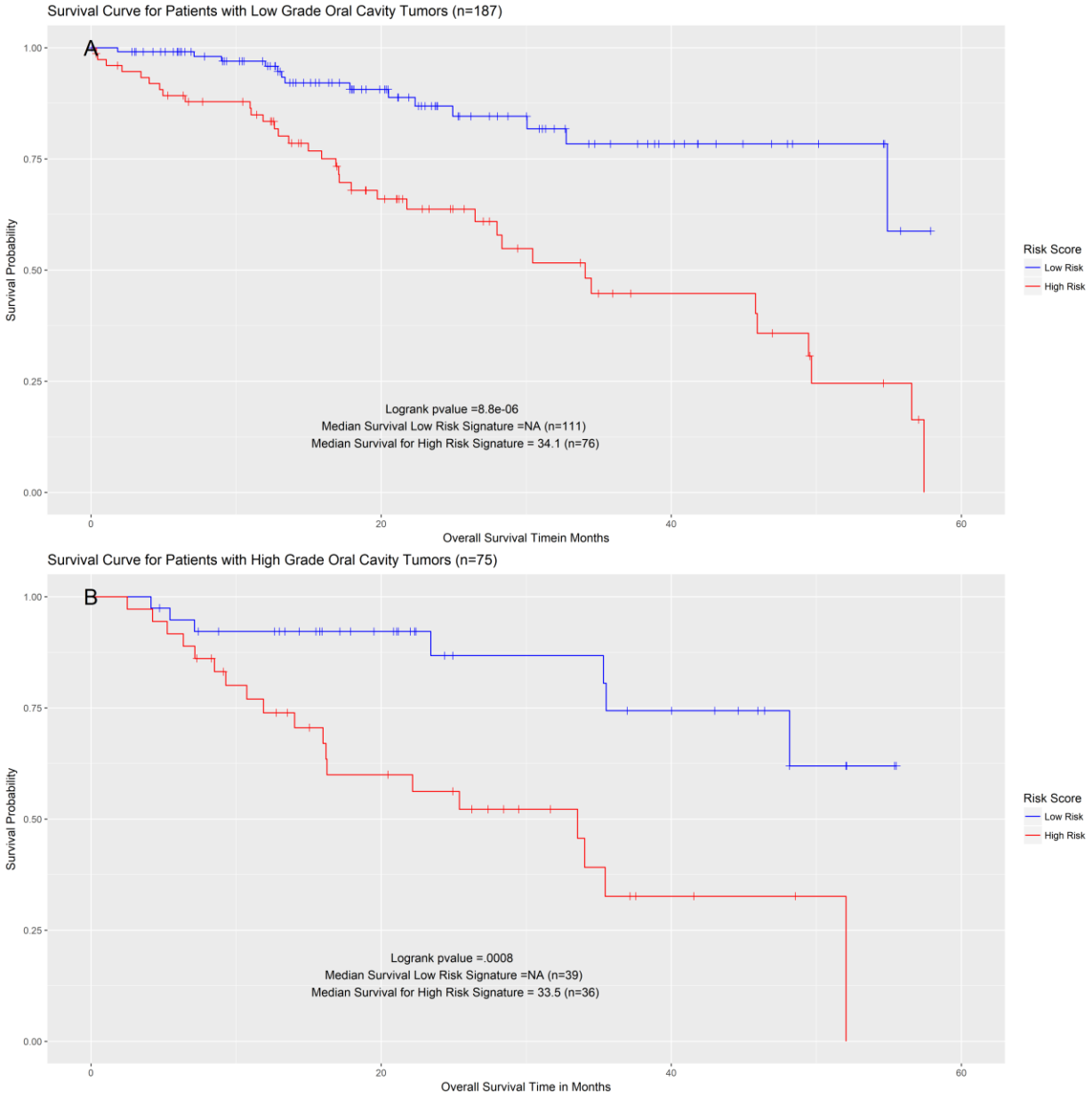
Table 3.3. Pathway analysis produced using Pathway Reactome. The "Fisher's exact p-value" represents the probability that the genes would be selected if they were selected by chance alone. Only pathways with a p-value less than .05 were listed in this table. The false discovery rate (FDR) was also calculated but not shown here. The FDR represents the probability that a gene is significantly enriched in error. The FDR is considered to be a conservative measure of significance, as it is not weighted to adjust for the number of times a gene was identified over 100 runs. Of the pathways listed only the first "Ligand-gated ion channel transport" had an FDR p-value of less than .05 (p-value=3.43E-04)

Figure 3.1. Validation of Aggregated Signature by Histogram, ROC Curve, Overall Survival Plot



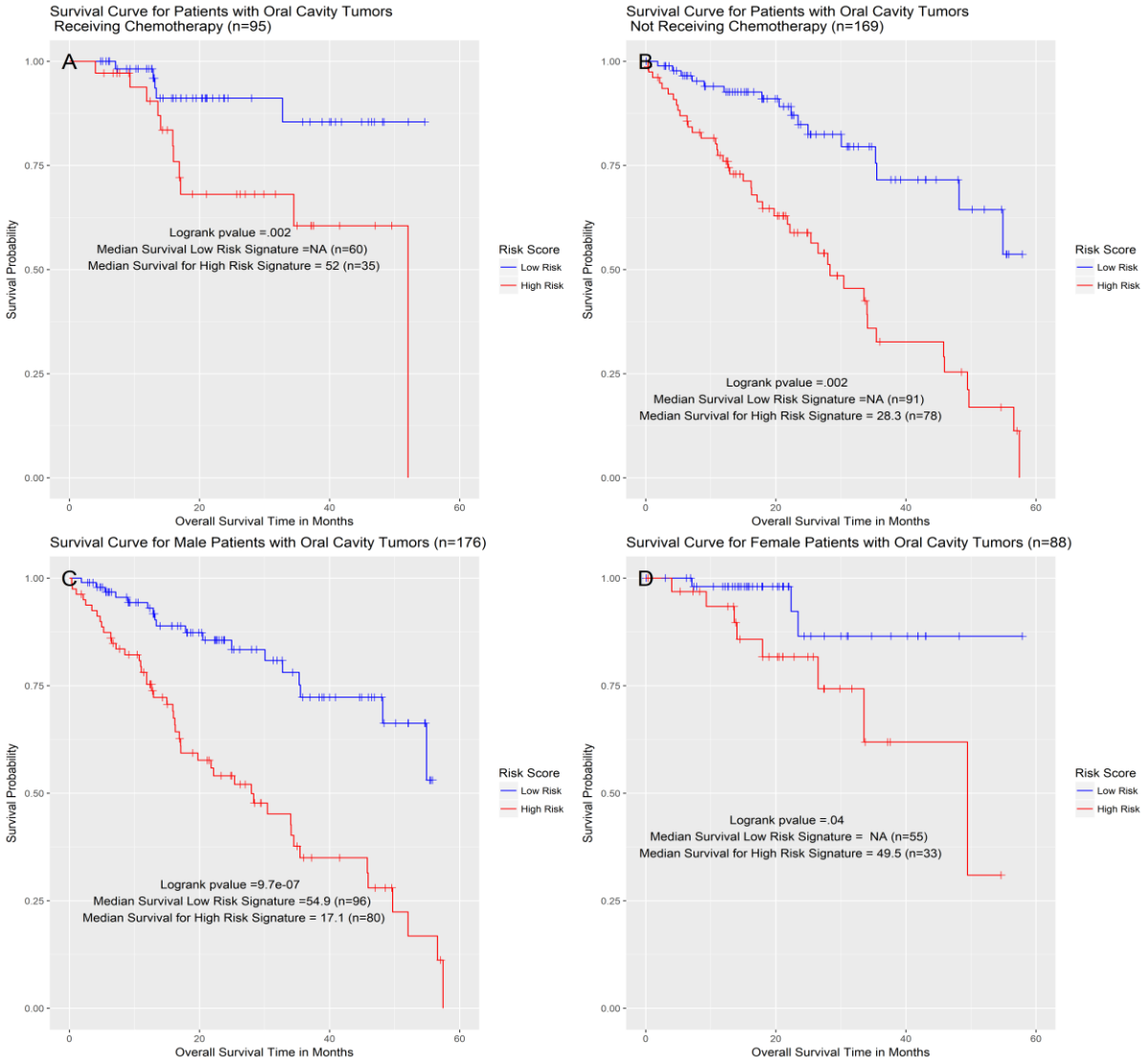
ROC curve threshold was selected by taking the point where there was a minimal difference between sensitivity and specificity. True Positive Fraction is synonymous with “Sensitivity”, False Positive Fraction is synonymous with 1-Specificity.

Figure 3.2. Survival Analysis, Stratifying by Tumor Grade



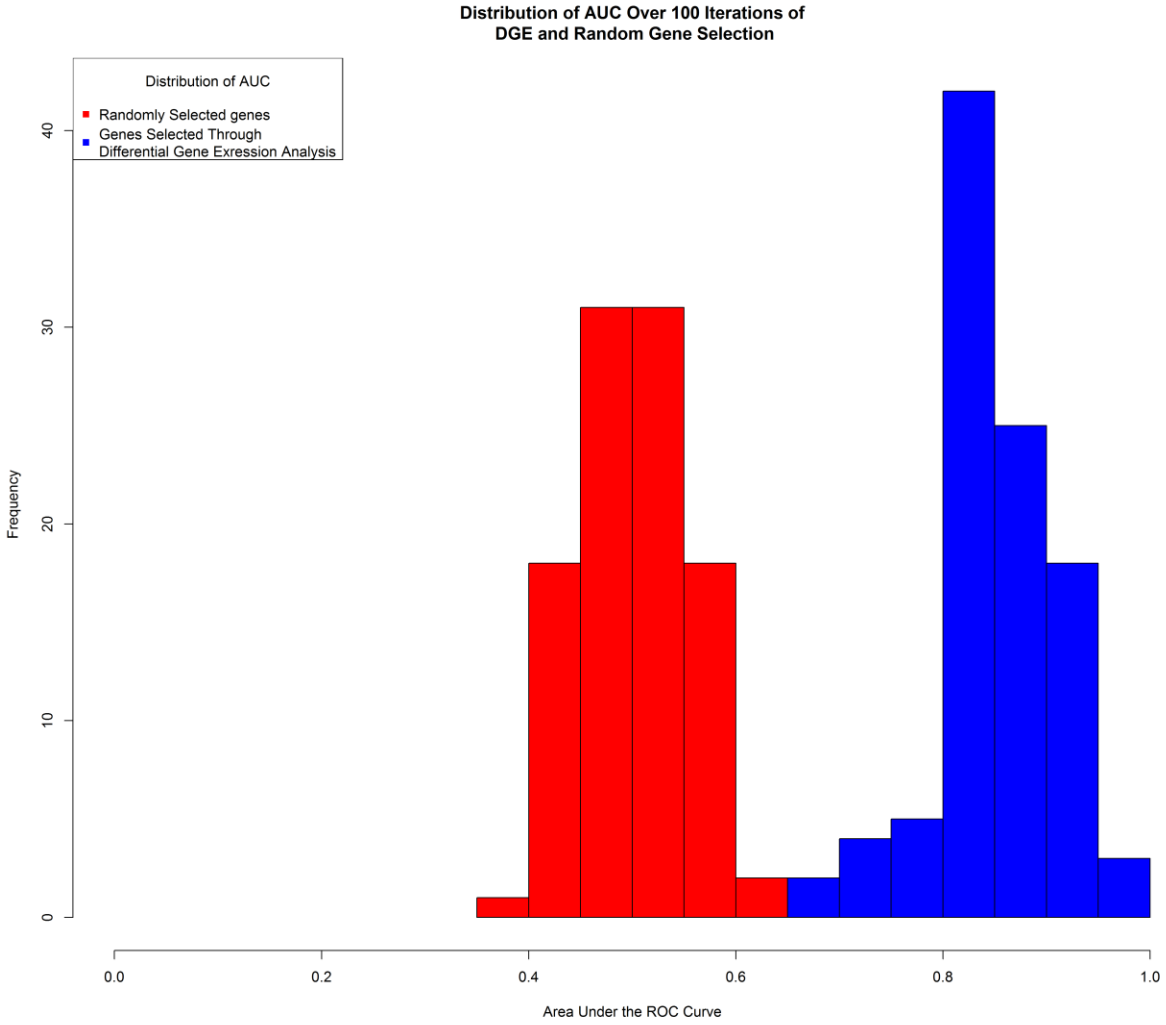
High tumor grade in the lot refers to patients with tumor grade of three or greater. Low tumor grade refers to patients with grade of grade 2 or lower.

Figure 3.3. Survival Analysis Stratifying by Chemotherapy Treatment Status, Survival Analysis Stratifying by Gender.



Patient chemotherapy status was divided into: “received any type of chemotherapy”, “did not receive any type of chemotherapy”.

Figure 3.4. Distribution of AUC of 100 signatures selected via DGE and Distribution of AUC of 100 Signatures using Randomly Selected Genes.



This figure shows that the distribution of AUC produced by 100 runs of DGE did not overlap with the distribution of AUCs produced from gene signature produced from randomly selescted genes.

Table 3.4. Frequency of Gene identification over 100 runs of Cross Validation

Gene Name	Frequency Identified over 100 iterations	Full Gene Name	Cox Model Variable Coefficient in Aggregated Signature.
DHRS7C	34	Dehydrogenase/reductase (SDR family) member 7C	0.001
CDX2	30	Caudal type homeobox 2	0.004
RETN	30	Resistin	0.001
FGF21	28	Fibroblast growth factor 21	-0.003
ANKS4B	27	Ankyrin repeat and sterile alpha motif domain containing 4B	0.009
RNF17	26	Ring finger protein 17	0.0003
SERPINA12	21	Serpin peptidase inhibitor	0.001
LEFTY2	20	Left-right determination factor 2	-0.006
LOC158696	20	LOC158696	-0.005
WDR87	20	WD repeat domain 87	0.001
KRT72	20	Keratin 72;	-0.03
GLP2R	19	Glucagon-like peptide 2 receptor	0.1

EXD1	19	Exonuclease 3'-5' domain containing 1	0.06
ITIH1	18	Inter-alpha-trypsin inhibitor heavy chain 1	0.004
SP7	18	Sp7 transcription factor	0.01
GLRA4	17	Glycine receptor, alpha 4	-0.01
KLK3	17	Kallikrein-related peptidase 3	0.04
C1orf194	17	Chromosome 1 open reading frame 194	-0.05
GDF7	17	Growth differentiation factor 7	0.0001
HTN3	16	Histatin 3	-0.003
AKR1D1	16	Aldo-keto reductase family 1	0.0004
LOC100192378	16	LOC100192378	-0.0006
C20orf56	16	OCSTAMP	0.0005
GP2	16	Glycoprotein 2 (zymogen granule membrane)	0.03
SLC22A25	16	Solute carrier family 22, member 25	-0.06
FAM138F	16	Family With Sequence Similarity 138 Member F	-0.009

CASR	15	Calcium Sensing Receptor	-0.04
ACSBG2	15	Acyl-CoA Synthetase Bubblegum Family Member 2	-0.15
HERC2P4	14	Hect Domain And RLD 2 Pseudogene 4	-0.007
C20orf123	14	Interleukin 1 family, member 10	0.01
IL17A	14	Interleukin 17A	-0.007
C10orf107	14	Chromosome 10 open reading frame 107	0.007
ASB4	14	Ankyrin repeat and SOCS box containing 4	-0.0004
CCDC129	14	Coiled-coil domain containing 129	-0.01
CALCR	13	Calcitonin receptor	-0.0007
HTR3C	13	5-hydroxytryptamine (serotonin) receptor 3C, ionotropic	4.74E-05
DCT	13	Dopachrome tautomerase	0.0004
KRT38	13	Keratin 38	0.04
SFTPC	13	Surfactant protein C	0.04

Risk score= $\log[\sum_{i=1}^{40}(c_i * cpm_i)]$, cpm=Gene Expression Counts per Million, c_i= "Cox Model Variable Coefficient in Aggregated Signature for gene i".

Table 3.5.Number of Literature Citations Detected For Each Gene in Final Aggregated Signature

Pubmatrix Terms	Oral Cancer	Head and Neck Cancer	Chemotherapy Resistance	Treatment Response	Overall Survival	Cancer Survival	Tumor Progression	Tumor Metastases	Tumor Necrosis
KRT72(keratin 72, K6IRS2, KRT6IRS2,KRT6,K6irs*)	34	82	2	37	116	103	40	92	23
DHRS7C	1	0	0	0	1	0	0	0	0
CDX2	10	99	8	51	214	185	144	187	29
RETN	1	2	41	47	33	4	13	6	127
FGF21	0	1	67	125	49	8	9	1	35
ANKS4B(FLJ38819, HARP*)	4	8	7	52	142	51	37	23	21
RNF17	0	0	0	0	0	0	0	0	0
SERPINA12	0	1	17	8	5	0	1	0	13
LEFTY2(LEFTY,LEFTA,LEFTYA, EBAF,TGFB4*)	0	1	1	13	19	6	7	3	6
LOC158696	0	0	0	0	0	0	0	0	0
WDR87	0	0	0	0	0	0	0	0	0
GLP2R	0	0	2	4	3	0	0	0	1
EXD1	0	0	0	0	0	0	0	0	0
ITIH1(H1P,IATIH,ITIH,IGHEP1,ITI-HC1)	0	0	0	1	3	2	2	1	1
SHAP	0	0	0	4	9	7	3	2	4
SP7	2	0	2	27	47	5	3	1	21
GLRA4	0	0	0	0	0	0	0	0	0
KLK3(PKK8,PKKD,PPK*)	14	15	8	42	72	33	47	20	3
C1orf194	0	0	0	0	1	1	0	0	0
GDF7	1	3	0	4	1	1	2	0	0
HTN3	1	1	8	7	14	0	1	0	3
AKR1D1	0	0	1	1	1	0	0	0	1
LOC100192378	0	0	0	0	0	0	0	0	0
C20orf56	0	0	0	0	0	0	0	0	0
GP2(ZAP75*)	0	1	1	27	30	10	4	9	11
SLC22A25(UST6, HIMTP,MGC120420*)	0	0	0	0	0	0	0	0	0
FAM138F	0	0	0	0	0	0	0	0	0
CASR (FHH,NSHPT,GPRC2A,CAR,EIG8 ,FIH,PCAR1*)	6	196	150	1302	2457	622	179	127	207
ACSBG2	0	0	0	0	0	0	0	0	0
HERC2P4	0	0	0	0	0	0	0	0	0
C20orf123	0	0	0	0	0	0	0	0	0
IL17A	7	10	12	186	112	32	51	11	222

C10orf107	0	0	0	0	1	0	0	0	0
ASB4	0	0	0	0	1	0	0	0	2
CCDC129	0	0	0	0	0	0	0	0	0
CALCR	0	3	0	4	7	2	1	1	5
HTR3C	1	1	1	4	0	0	0	1	0
DCT	10	15	11	84	90	45	23	20	14
KRT38	0	0	0	0	0	0	0	0	0
SFTPC(SP-C,PSP-C,SMDP2,BRICD6)*	117	181	327	1460	2072	452	194	176	483

All genes listed within parentheses preceding an asterisk were known aliases of the preceding gene which was included within the aggregated signature. Aliases were included to provide greater ability to detect citations for genes included within the final aggregated signature.

Chapter 4

Identification of Targetable Pathways in Oral Cancer Patients via Random Forest and Chemical Informatics

Introduction:

Treatment of head and neck cancer has also been slow to change with EGFR inhibitors, PD1 inhibitors, and taxol based chemotherapies being the only therapies approved by the FDA in the last 10 years for head and neck cancers. Head and neck cancer is relatively rare compared to more common cancers like breast or lung cancers. However, it is possible that existing therapies for more common solid tumors could also prove effective in oral cancers. Many therapies have molecular targets that could be appropriate in oral cancer as well as the cancer in which the drug gained initial FDA approval. There may also be targets in oral cancer that may be viable for which existing FDA approved drugs could be applied. This study describes informatics methods that utilize machine learning to identify influential gene targets in patients receiving platinum based chemotherapy, non-platinum based chemotherapy, and genes influential in both groups of patients.

Drugs approved by the FDA for oral cancer are methotrexate, cetuximab, pembrolizumab, nivolumab, and docetaxel. These therapies are combined to be used in conjunction with platinum based chemotherapies such as cisplatin or carboplatin unless those therapies are contraindicated due to comorbidities such as renal disease. The small number of new oral cancer drugs could be attributed in part to the low overall burden of oral cancer in comparison with other cancers. The current timeline for FDA

approval of a novel small molecule or biologic is 10 years or more. Using existing FDA approved drugs is a popular method utilized to shorten this process to 3-4 years. Cetuximab is an EGFR inhibitor that has been shown to decrease the rate of progression of oral cancer used in conjunction with cisplatin. Pembrolizumab and Nivolumab are both PD1 inhibitors that utilize T-cells to attack cancer while it progresses. These therapies use the bodies immune system as another treatment modality to reduce the burden of oral cancer. Current literature provides support for the role of ligand channel gating, hedgehog signaling[6–8], NOTCH, B-WICH[9], inflammasome[10], WNT[11], and Calcineurin pathways in cancer[12–14]. The role of these pathways and targeting specific genes within them has been pursued in other cancers, but have with the exception of only a few studies, not yet been examined in oral cancer. Possible gains from targeting these pathways would be initiating immune response, targeting cancer metabolism, targeting signaling for metastasis, and targeting inflammation pathways that may drive progression. If there is a synergistic effect to attacking multiple hallmarks of cancer simultaneously then the net gain to the patient would be in overall post diagnosis survival time for the oral cancer patient.

Identifying a means by which drugs may be prioritized for further screening and validation for a specific cancer type would be desirable. Databases linking genes, the proteins they express, ligands corresponding to those proteins, and structural data that can be analyzed all exist in varying forms or completeness across different publicly available databases. This study describes how integration of analyses of these databases can be used to select gene targets in a specific cancer, and how therapies

can be prioritized for screening based upon existing structural information for the ligands associated with genes and the proteins they express.

There are several hurdles to the analysis of high dimensional genomic data using traditional regression analyses. Random forest analysis is a machine learning approach that is less hindered by datasets with large predictor to observation ratios. In this study we apply random forests analysis to gene expression data to identify those genes, and pathways that are most predictive of post diagnosis survival across treatment strata. This is the first application of this approach to head and neck cancer patient data in the cancer genome atlas (TCGA). National Comprehensive Cancer Network (NCCN) guidelines recommend that node positive patients with tumors of clinical stage 3 and greater receive chemotherapy[15]. Platinum based chemotherapy with radiation and surgery is the current standard of care recommended to these patients. Patients that do not receive chemotherapy recommendations by NCCN are node negative with clinical stage 2 and lower. Following standards set by NCCN guidance on treatment, this study chose to identify patients receiving and not receiving platinum based chemotherapy as separate groups. Analysis of influential genes in each group will improve knowledge of possible mechanisms driving treatment response for early stage and more advanced tumors.

Random Forest is a machine learning approach to identifying the most important predictors in high dimensional datasets[16–18]. This approach is uniquely suited for classification of observations in datasets where P (predictors) are > than N (number of observations). Random forest randomly selects predictors from a large group of

predictors and then applies those predictors to a decision tree predicting overall survival. Random forest does not pay a statistical penalty when the number of observations is small. Instead the strength and limitation of this method is its reliance upon computational intensity. That is, as the number of decision trees in a random forest increase so does classification accuracy. Accuracy is also dependent upon the number of predictors tried at decision tree nodes. As node size and forest size increase so does forest classification accuracy. However, there is a rate of diminishing returns in the accuracy gained from each tree added to a forest. This is why computational time and cost must be factored into all random forest analysis plans to measure project feasibility. Random Forest has been successfully applied to predicting cancer diagnosis, and treatment response for a variety of cancers[19–23]. For this study we have selected to apply random forest analysis to the gene expression values of oral cavity cancer patients to identify the upregulated pathways most predictive of improved treatment response across gender and environmental exposure subgroups like alcohol and tobacco. RNAseq data is inherently high dimensional, applying typical regression models to such data can be costly as large sample sizes are required to identify even moderate effect. Identifying gene interactions can be even more costly in terms of the required statistical power. Stratified pathway analysis via random forest methods has been shown to be successful in identifying single influential genes (within the context of larger pathways) that are predictive of overall survival with limited sample size[24]. This approach has not yet been applied to identification of influential genes and gene interactions within oral cancer patients stratified specifically by treatment. In this way the importance of pathways and genes of interest can be compared across strata to

assess which subgroups may be most sensitized to changes in gene expression within a given pathway.

Methods

This study focuses upon the identification of the role of gene expression in oral cavity cancer patients and applying machine learning approaches like random forest to determine genes that are important in influencing treatment response. Reference ligands known to bind to proteins expressed by genes deemed influential by random forest, can be sent through a virtual screening pipeline to identify small molecules with greater likelihood of acting as protein agonists/antagonists. Ligands that have a strong shape similarity to known binding ligands have greater potential for success in high throughput screening endeavors. Of course shape similarity alone is insufficient in identifying new drug leads, this is why all leads will also be validated with existing literature, and those leads without previous biological validation will be presented as such.

By using a stratified random forest analysis we will be able to rank genes within the strata of chemotherapy treatment status. This approach will allow for the identification of those top ranked genes that are unique to each stratum. This will be done by identifying common and unique genes between sets of genes influencing the treatment response in patients receiving platinum based chemotherapy and those that do not. The end result will be the identification of oral cavity cancer pathways influencing treatment response which will inform researchers on mechanisms driving treatment response in

specific groups such as those late stage, node positive patients that are more likely to receive chemotherapy treatment. This analysis will illustrate and support existing studies showing the strength of machine learning methods as an alternative method in identifying gene expression values influencing treatment response. This study is focused not on the predictive power of an aggregated panel of gene expression values, but rather to integrate random forest with chemical informatics and thus describe methods to shorten the pursuit of novel therapies treating cancers with relatively lower incidence.

Retrieval of Public Data

This study used clinical and genetic data obtained from The Cancer Genome Atlas (TCGA). Genetic data included raw counts per million (CPM) of RNA sequence expression values for 523 patients posted to TCGA. Of these 523 patients, 313 were diagnosed with Oral Squamous Cell Carcinoma (OSCC). Oral squamous cell carcinoma patients included tongue, buccal mucosa, alveolar ridge, general oral cavity, and soft palate tissues. Of these 313 patients, 267 were included based upon complete survival time. Of these 267, 109 received either Carboplatin, Oxaliplatin, or Cisplatin while 158 patients received a treatment other than the platinum based chemotherapy treatment regimen. All tissue samples were collected prior to start of treatment. Clinical data on tumor stage, necrosis, size, and nuclei were also retrieved from TCGA. We obtained demographic data on ethnicity, race, and gender from TCGA clinical files. In addition, this dataset had information on environmental exposures like tobacco history (ever/ never smoke), number of alcohol containing drinks in a day (greater than two drinks consumed per day, two drinks or less consumed per day), Overall survival time in months was extracted as a measure of treatment response.

Machine Learning Methods

Description of Random Forest Approach

Stratified pathway analysis considers important covariates in data analysis. In this analysis of head and neck cancer gene expression data, this study used information regarding patient age, gender, alcohol and smoking exposures. In the first round of analysis the sample of 267 oral squamous cell carcinoma patients were organized into 109 and 158 subsamples based upon whether patients received platinum or non-platinum based chemotherapy respectively. For each group, this study built a random forest[18] to predict survival time based upon the gene expression levels within each subgroup.

To better communicate the function of random forests, understanding of a decision tree construction is needed. A decision tree is constructed by:

Step A: Taking a bootstrap sample from the original sample.

Step B: A decision tree is grown for each bootstrap sample.

Step C: At each tree node apply a predetermined number of predictors randomly to create branches within the tree.

Step D: A branch is formed using the predictor from step C.

Step E: Repeat steps C and D until the end of every tree branch contains samples above or below the same survival threshold or contain only one sample.

Random Forest Result Measurements

Random Forests build many decision trees to comprise a forest. Each tree is put together by using a random bootstrap sample of the original data and applying a random number of predictors at each node of a decision tree. The SRCRandomForest R package [25] employed in this study sets aside half of the data to be used for validation purposes to measure the accuracy of the random forest model. The p-values yielded through this analysis are defined as the "proportion of cross validation errors smaller than the cross validation errors obtained from 500 iterations of random forest runs of randomly permuted labels of patients". This list of genes can be used to identify pathways that are enriched with the influential genes identified through random forest at odds that would be greater than can be attributed to chance alone with a p-value=.05. This analysis will present pathways common to, and unique to, each chemotherapy treatment strata. Such analyses may identify plausible biological mechanisms that enhance understanding of observed differences in survival. To reiterate, the focus of this study is not to pursue a diagnostic tool but to identify those gene expression values exerting a strong influence on treatment response. This study adopted random forest as the machine learning method of choice due to its superior interpretability and scalability.

Random Forest Tuning Parameters

Our random forest model used 20,000 tree forests for a forest size, with 320 variables tried at each node in each decision tree. 20,000 trees was the point at which we could identify no significant increase in our ability to predict patient survival. 320 tries at each node in each decision tree was double the recommended number of tries given by the author of our R software package RandomForestSRC. The author H. Ishwaran refers to a generally accepted practice of using the "square root of the total number of predictors as a starting point for the number of variables tried at each node"[25]. It was for this reason that we applied 20000 trees and 320 variables tried at each node in each decision tree for every strata in our analysis. This approach was applied to our entire final sample of 2677 OSCC patients. We then divided this sample by whether a patient received platinum based chemotherapy or not. The output of each group's analysis produces a list ranking each gene. This analysis identified common and unique pathways between the entire dataset and each chemotherapy treatment group. We then identified the unique and common genes between chemotherapy groups. This analysis will allow us to observe the difference in gene importance and corresponding pathways in relation to overall survival.

Description of Virtual Screening Approaches

Using chemical informatics techniques, ZINC drug database [26] of 1,379 FDA approved drugs (FDA), and ZINC Traditional Chinese Medicine database of 39,894 small molecules (TCM) can be used to apply three-dimensional chemical informatics approaches to the identification of small molecules that are the best candidates for inhibition of proteins expressed by those genes influencing treatment response. Reference ligands for each protein are obtained from the Royal Chemistry society Protein Database[27] and then virtually screened against FDA and TCM small molecule libraries. Molecular shape overlay is an approach for measurement of the similarity of one molecule in comparison with another. A Tanimoto coefficientis used to measure the degree of similarity between two molecules. A goal of this study is to perform searches of two small molecule databases FDA and TCM using a maximum common substructure measurement of Tanimoto similarity from the R Rcpi package[29] that has been shown to perform robustly across a variety of molecule types.

The Tanimoto coefficient between two points, *a* and *b*, with *k* dimensions is calculated as:

$$\frac{\sum_{j=1}^{k} a_j \times b_j}{\left(\sum_{j=1}^{k} {a_j}^2 + \sum_{j=1}^{k} {b_j}^2 - \sum_{j=1}^{k} a_j \times b_j\right)}$$

The Tanimoto similarity only applies to binary variables, for binary variables the Tanimoto coefficient ranges from 0 to 1 (where 1 is the highest possible similarity).[28]

Pathway Analysis

Table 1 provides information on those genes that are significantly enriched within our gene set beyond what would be expected by chance alone. The significance of enrichment is calculated as the odds of randomly selecting the number of genes in the submitted set of genes by randomly selecting from 20,530 genes over 100 times. The

false discovery rate (FDR) reported by the Pathway Reactome application utilized for this study represents an adjustment for multiple comparisons across all pathways. Benjamini-Hochberg false discovery rate is calculated as the p-value ranking (smallest being 1 and all following having greater or equal rank dependent upon size of p-value) divided by the number of tests performed, multiplied by the significance criterion. In this study .05 is the criterion used to measure significance. The false discovery rate can be interpreted as the proportion of tests within a set of tests that falsely rejects the null hypothesis. If a FDR. Is .5 then 50% of those pathways identified falsely reject the null hypothesis. It is important to note that the FDR calculation utilized by Pathway Reactome defaults to perform a large number of analyses/tests. Additionally there are many pathways examining similar genes and gene types. Unfortunately this thorough examination strategy also inflates the number of analyses and causes the FDR to become overly conservative.

To identify those genes that are most likely to be connected to influential pathways two filters were applied to gene selection. First the gene had to be in the top 5 of influential genes identified via random forest. Second, if the gene was not within then to 5% of genes it could still be included within the analysis if it was within the to 40% of genes and was known to be connected via past experimental studies supporting the Cytoscape/Pathway Reactome plugin database. This can be accessed by uploading topology from a given gene and merging it with that gene's corresponding random forest importance values. Thus, the selection of genes included those ranked as the most important by random forest (top 5%) or of moderate importance and high topology

(shown through past experiment or literature search to be connected to 50 genes or more). For each set of analyses 1000 influential genes were selected and 100 high topology genes were selected. In this way, integration of network analysis with random forest allowed for identification of pathways significantly enriched with the genes identified by this approach.

Results

Top Pathways

This section will describe pathways uniquely influential in patient response to platinum based chemotherapy, and influential in response to non-platinum based therapy. Influential pathways shared by both platinum based chemotherapy users and non-users will also be presented. An influential pathway will be defined as a pathway that is significantly enriched with genes that were in the list of top 1000 (5%) of most influential genes yielded by random forest analysis for platinum based chemotherapy users, non-platinum based chemotherapy users, and for those pathways enriched with genes shared in common in lists of top 1000 genes for platinum based chemotherapy users and non-users. The top 5% of genes were conservatively selected to produce greater certainty of the link between highly ranked genes, the pathways in which they were enriched, and the link between treatment response and significant pathways identified through gene enrichment analysis. Top pathways could also be enriched with those genes in the top 40% of important genes identified by random forest if they also had a high amount of connectedness (a gene was connected to 25 gene nodes in a gene network) reported by the Cytoscape/reactome application.

Pathways significantly enriched with genes identified as the most influential (top 5%) in predicting treatment response for platinum based chemotherapy users were those related to calcium channel gating, hedgehog signaling, histone acetylation, elastic fiber production, tRNA acetylation, hexokinase deficiency, inhibition of adenylate cyclase, CLEC7A inflammasome. It has been reported that calcium channel gating has been associated with multiple cancers[30–32]. There have also been recent studies evaluating the benefit of targeting histone deacetylation pathways in oral cancer[33–35]. The hedgehog signaling pathway has also been shown to signal progression in other cancers “Hh signaling has been shown to regulate the self-renewal of CSCs in breast, glioma and multiple myeloma, and more convincingly in the maintenance of chronic myelogenous leukemia (CML) stem cells”[36–41]].” All significant pathways for platinum based chemotherapy users are in Table 1.

Significant pathways for patients not using platinum based chemotherapy were those related to B-WICH complex,TP53 pathway, FGFR pathway, potassium channel gating, and RNA polymerase chain elongation pathways and their epigenetic regulation. TP53 and FGFR pathways represent the expression of canonical oncogenes which have been shown to be cancer drivers and associated with the production of all cancers.[42][43–45].The B-WICH complex has been found to be linked to maturation of invadopodium in breast cancer and has been suggested as both a biomarker and target for cancer invasiveness [46,9]. Potassium channel gating has also shown to be a potential target for head and neck cancers due to this pathways association with

immune response and treatment response [47–51]. RNA polymerase chain elongation and its role in transcription is a logical contributor to cancer progression and differentiation, however the lack of specificity makes this a difficult pathway to target specifically in cancer cells. All significant pathways for non-platinum based chemotherapy users are in Table 2.

For those pathways enriched with genes shared by users and non-users of platinum based chemotherapy it was found that pathways related to g-protein beta folding, NFAT activation and repression of Wnt pathway genes were enriched with genes influential in treatment response in both groups of patients. Repression of Wnt pathway genes may be done through targeting the sonic hedgehog pathway as previously outlined or through more direct means which have been researched in multiple other cancers.[52–54]. NFAT proteins have been found to be associated with cancer progression in blood and solid tumors however the literature is mixed as to whether NFAT pathways are viable targets for treatment. [55–57] These pathways influencing treatment response in both users and non-users of platinum based chemotherapy can be seen in Table 3.

Important genes and biological implications

A visualization of pathways overlapping between users and non-users of platinum based chemotherapy highlight the importance of several genes in a way that random forest analysis alone could not. By visualizing the four common pathways it becomes possible to identify not only highly influential genes but those genes that have the highest degree of connectivity to influential genes. Using annotation built into

Cytoscape[58,59] we can also identify existing small molecules used in cancer therapy that are not yet commonly used in oral cancer, and we can also observe those genes previously found to be associated with oral cancer. Genes found to be influential in oral cancer for patients receiving platinum based chemotherapy with existing literature supporting the targeting of these genes in cancer were INSR, BRAF, and PSMB7 which are targeted by ceritinib, (regorafenib&dabrafenib), and bortezomib respectively. These drugs are not currently FDA approved for treatment in oral cancer. Genes found to be influential in oral cancer for patients not receiving platinum based chemotherapy with existing FDA approved chemotherapy drugs targeting the products of said genes are FGFR3, EGFR, PRKAA2, CSNK2A1, INSR, MET, CAMK2A, PSMB5, and PSMB1.

There are multiple chemotherapy drugs targeting these pathways with 14 different drugs targeting EGFR alone. It should be noted that EGFR is a gene pathway being targeted in current oral cancer treatment. Sarafenib Tosylate, Pazopanib Hydrochloride, and Vadetanib all target the FGFR pathway specifically. Sunitinib Malate is unique in that it has been found to act on 4 different genes that were found by random forest to be influential in treatment response FGFR3, CDNK2A1, PRKAA2, and CAMK2A. Again we see that Certinib acts on a gene that is influential in both patient treatment groups, this gene is INSR. Bortezomib, Carfilzobib, and Ixazomib all act on PSMB5 which is an influential gene in both Platinum and non- platinum based therapy. PSMB5 and PSMB1 are both found to be within the top 5% of influential genes in random forest analysis and are genes that are significantly enriched within the sonic hedgehog pathway. Chemotherapy drugs and their relationship to genes in common influential pathways

between users and non-users of platinum based chemotherapy are visualized in Figure 1 for platinum based Chemotherapy users and Figure2 for non-platinum based chemotherapy users.

In addition to analysis of the intersection of existing cancer drugs and genes deemed influential by random forest, this study also looked at the intersection between gene topology within a pathway and random forest influence. This analysis identified CTNNB1, PLCG2, SHC1, UBA52, UBB, UBC, and HDAC3 as genes that meet filters of belonging to one of the four common enriched pathways, being a gene that is one of the top 5% of influential genes listed by random forest analysis, and being connected to over 100 genes within the four interconnected pathways (Figure 1, Figure 2). CTNNB1 mutations have found to be predictive of Lung and other thoracic cancers [60–64], PLCG2 and calmodulin knockdown have been shown to induce paclitaxel sensitivity in cervical cancer tumors. This may prove of use to oral cancer patients which may be assigned to paclitaxel or other taxol regimen [65,66]. SHC1 has been shown to be a regulator of EGFR function and thus a potential target for multiple cancer types where EGFR is a key driver [67–69]. Ubiquitin genes UBA52, UBB, and UBC have been shown to be associated with several cancers and research is currently being pursued in targeting ubiquitin ligases to improve treatment response [70–74]. Histone-Deacetylase genes specifically HDAC3 is shown to be a hub to several genes that are influential in platinum based chemotherapy response genes that have been associated with metastatic invasion in breast and pancreatic cancer Figure 1). Inhibition of HDAC3 was shown to impact signaling to cancer stem cells This gene has been shown to be a

regulator of apoptosis control exerted by TP53 [75–78]. These genes were not shown to currently have any antibody or small molecule therapies targeting their action. High level topology (over 100 connected genes) and high random forest importance ranking should provide impetus for further research into the targeting of gene action in oral cancer.

Chemical Informatics Analysis of Drug Targets and Leads

For those genes meeting topology and random forest filters, the known ligands of proteins expressed by each gene were identified though the RCSB protein data bank. Structural files of ligands were downloaded as .sdf files and uploaded into the chemical informatics R package Rcpi. Once loaded, each ligand had to undergo virtual screening against all FDA approved drugs to identify existing FDA approved drugs that may prove efficacious as therapeutic agents. Only those molecules with a Tanimoto similarity score > 50% were included in results. A Traditional Chinese Medicine (TCM) small molecule database was also utilized as biologic derived small molecules are known to provide better shape overlay when screened against other biologic small molecules. Additionally the molecules in the TCM database have been shown to be generally safe in people by merit of its long historical use in human populations. For CTNNB1 several ligands were identified via RCSB PDB (2s)-3-{[{[(2s)-2,3-dihydroxypropyl]oxy}(hydroxy)phosphoryl]oxy}-2-[(6e)-hexadec-6-enoyloxy]propyl (8e)-octadec-8-enoate was the single ligand associated with CTNNB1 that was used for virtual screening against the FDA and TCM libraries. Unfortunately neither library yielded a small molecule candidate with greater than a 50% Tanimoto score.

The Ligand of PLCG2 did yield several interesting drug leads in both the FDA approved library and the TCM library. Fludarabine has been tested in oral cancer cell lines and found to be effective in inducing cell apoptosis[79]. Ganciclovir an HIV drug has also been tested in oral cancer and found to have a clinical effect on cell differentiation[80]. Entecovir and didanosine are drugs used in the treatment of HepB infection and HIV and have not yet been tested on oral cancer. Small molecules in the TCM database meeting shape overlay filters were [4-(2,6-dimethylmorpholin-4-yl)sulfonylphenyl]-[4-(2-phenoxyethyl)piperazin-1-yl]-methanone which has not yet been used on oral cancer cell lines. There was overlap between drug leads for ligands of UBB and PLCG2. This is due to the structural similarity between cytosine and guanine ligands used as reference molecules for similarity matching. Drug leads for UBB included Cytarabine (Cancer), Fludarabine (Cancer), Azacitidine (myelodysplastic syndrome), Gemcitabine (Cancer), and Lamivudine (HIV). Gemcitabine is unique in that it is the only drug of those listed, that has been approved by the FDA for use in oral cancer patients. For HDAC3 there were no matches exceeding a Tanimoto score threshold of 50% of the reference molecule when using a library of FDA approved drugs. The TCM database did yield a match with an extract from *Mallotus Phillipinensis* a member of the Euphorbiacae plant family. An extract of this plant known as Rottlerin has been found to inhibit growth of colon cancer, and breast cancer cells [81,82].

Quercetin, and Diosmetin were other phenols found in citrus that were also identified as matches meeting Tanimoto thresholds. There are no reports of the effect of quercetin or diosmetin in oral cancer. It should be noted that Quercetin, and Rottlerin have been noted in literature as promiscuous ligands that are often found in natural product insilico

screenings[83]. A recent study reported quercetin as the number one natural product in terms of number of occurrences within the database[84]. The aforementioned studies do show that Rottlerin is related to the metastatic potential and viability of colorectal cancer cells[81]. Caution and validation of results with existing literature or carefully designed follow up experiments should always be pursued to justify the results of promising insilico analyses. These analyses enhance knowledge of genes influencing treatment response in oral cancer. Pathway, network, and chemical informatics analysis can be paired with a literature review to identify drug leads for oral cancer treatment. Reference ligands associated with influential genes in the Royal Chemistry Society Protein Database are listed along with Drug leads and their corresponding Tanimoto similarity scores in Table 4.

Discussion

A machine learning approach known as random forest was used to identify genes influencing oral cancer treatment response specific to the platinum based chemotherapy treatment type, and the non-platinum based chemotherapy treatment type. This paper emphasizes the benefits on integrating the results of this line of analyses with pathway, network, and chemical informatics analysis to identify promising gene targets, and drug leads. Biological plausibility of these findings were highlighted with a review of existing literature supporting the findings for pathways, genes, and small molecules that our reported approach identified as influential in oral cancer. The results of this work identify pathways influencing treatment response in platinum based chemotherapy users, non-users, and those common to both users and non-users. Network analysis via Cytoscape allowed for the identification of those influential genes within each

treatment modality group within the context of inter connected gene networks. The utility of random forest was underscored in that in addition to pathways it also provides a rank to each gene in its influence on treatment response. This approach is a low cost method of prioritizing gene targets and drug leads. These methods are validated in that the genes identified have been shown to be associated with cancer progression in oral cancers and other cancers. Several drug leads identified were also shown to be effective in inhibiting oral cancer cells and were reported to be in different phases of the drug approval pipeline.

A possible criticism of the method outlined in this study is that there is uncertainty in the degree of trust that should be extended to random forest measures of gene influence, and the inference of importance to the pathways in which “influential” genes reside. To further such criticisms a point could be made that the Tanimoto threshold of >50% similarity could be perceived as low and the 50% difference in the molecules compared may prevent activity and may also be shown to have toxicity for a given disease state. Given such uncertainties it may seem that the evidence supporting these methods is tenuous. This study recognizes these criticisms, however, the counterpoints must be made that gene influence is not observed in a single sample of the data, but rather in over 20000 permuted samples of the data in which the top ranked genes were found to be more influential that thousands of other genes. The computational intensity provided in this study (20000 trees and 320 tries at each node of each decision tree) provides justification of the trust provided for each gene influence value. In respect to the results yielded by chemical informatics analysis it is important to note that the outlined chemical informatics method was able to identify gemcitabine a drug that has been approved for

use in oral cancer by the FDA. This method also identified Fludarabine, and Ganciclovir which have both been reported as providing significant reduction in oral cancer cell line progression and viability.

Drug leads were identified in both FDA and Traditional Chinese Medicine libraries, the benefit in expanding the number of libraries is that it increases the probability of finding a match meeting the Tanimoto threshold of >50%. The negative aspect of adding libraries is that if computational resources are not planned for accordingly then the amount of time required to screen against each reference molecule will scale upward with library size. The tools used in this study were all open source and freely available, a limitation to the adoption of this pipeline is that tools and their dependencies are distributed across different R repositories that may or may not be kept up to date. Combining these tools into a single package that allows for the identification of both gene targets and drug leads may enhance the pace of drug discovery pipelines. We have shown in this study that random forest is well suited to datasets with small observations and high number of features. Gene targets that have been shown (through literature review) to be associated with treatment response and cancer progression were identified through this study's use of random forest analysis. Stratifying this analysis by the type of chemotherapy received allows for interpretation of influential genes and pathways within the context of treatment. Indeed, the lack of overlap in the importance of genes from one treatment modality to another highlights that the gene expression patterns influencing platinum based treatment response differ from those gene expression patterns influencing non-platinum treatment response.

It is likely that there is bias inherent to the stratification of patients by chemotherapy treatment type. Chemotherapy treatment is associated with clinical variables like clinical stage, tumor size, and tumor grade, as well as gender and socioeconomic quintile[85–87]. This study attempted to address these confounders by including them within the bag of randomly selected features available for construction of the random forest model. By integrating chemical informatics analyses, random forest results can be translated into lists of drug leads for each target gene. This method identified drug leads that have already entered or passed phase three trials. Our review of identified drug leads and comparison with existing annotations show that the chemical informatics methods described can identify small molecules with therapeutic potential. This study provides the impetus for further exploration of the role of the identified small molecules in oral cancer treatment response, and the targeting of those genes identified as most influential by our series of analyses. This study also serves as a model for researchers identifying gene targets in rarer cancers where the number of cases is limited.

Table 4.1. Top Pathways Enriched with Genes Influencing Platinum Based Chemotherapy Treatment Response in Oral Cancer

Table 1. Top Pathways Enriched with Genes Influencing Platinum Based Chemotherapy Treatment Response in Oral Cancer			
Pathway name	p-value	FDR	Influential Genes Enriched in Pathway
Signaling by Hedgehog	6.2E-05	0.002	ARRB1;ARRB2;KIF7;ADCY6;PSMA7;ADCY5;PSMB6;TUBB6; PSMC6;PSME4;PSME1;PSME2;CDON
Molecules associated with elastic fibres	3.4E-04	0.006	ELN;FN1;FBLN1;LTBP3;BMP7
CLEC7A/inflammasome pathway	9.9E-04	0.01	IL1B,UBE2D4;ITPR2;ITPR3;PSMA7;PSMB6;PSMC6;IL1B; PSME4;PSME1;PSME2;TAB2;IKBKG;CALM1;CARD11

Phase 0 - rapid depolarisation	0.001	0.01	CAMK2B;CAMK2D;CACNB3;CAMK2A;CACNA2D2;CALM1
Adrenaline,noradrenaline inhibits insulin secretion	0.001	0.01	CACNB3;GNG2;CACNA2D2;ADCY6;ADCY5
Signaling by NOTCH1 in Cancer	0.006	0.03	HDAC5;HDAC1;EP300;CCNC;TBL1X
LGI-ADAM interactions	0.04	0.08	LGI2;ADAM11
SeMet incorporation into proteins	0.05	0.1	QARS
Presynaptic depolarization and calcium channel opening	0.05	0.1	CACNB3;CACNA2D2

Table 4.2. Top Pathways Enriched with Genes Influencing Treatment Response in Oral Cancer Patients not Receiving Platinum Based Chemotherapy

Pathway identifier	Pathway name	p-value	FDR	Influential Genes Enriched in Pathway
R-HSA-5250924	B-WICH complex positively regulates rRNA expression	3.2E-12	1.7E-10	HIST1H2BM;H2AFJ;H2AFZ;HIST1H2AJ; HIST1H2BK;H3F3A;POLR1C;H2AFV; HIST2H3C;HIST2H2BE
R-HSA-5250913	Positive epigenetic regulation of rRNA expression	8.8E-11	3.0E-09	HIST1H2BM;H2AFJ;H2AFZ;HIST1H2AJ; HIST1H2BK;H3F3A;POLR1C;H2AFV; HIST2H3C;HIST2H2BE
R-HSA-1296065	Inwardly rectifying K+ channels	2.4E-03	9.6E-03	GNG2;KCNJ14;GNB3
R-HSA-1839130	Signaling by activated point mutants of FGFR3	2.7E-03	1.1E-02	FGFR3
R-HSA-5655332	Signaling by FGFR3 in disease	3.9E-03	1.2E-02	KRAS;FGFR3
R-HSA-8853338	Signaling by FGFR3	3.9E-03	1.2E-02	KRAS;FGFR3

	point mutants in cancer			
R-HSA-2033514	FGFR3 mutant receptor activation	8.0E-03	2.4E-02	FGFR3
R-HSA-5654227	Phospholipase C-mediated cascade; FGFR3	3.5E-02	7.0E-02	FGFR3
R-HSA-6803211	TP53 Regulates Transcription of Death Receptors and Ligands	3.5E-02	7.0E-02	TNFRSF10D
R-HSA-2033515	t(4;14) translocations of FGFR3	4.7E-02	9.5E-02	FGFR3
R-HSA-5619109	Defective SLC6A2 causes orthostatic intolerance (OI)	4.7E-02	9.5E-02	SLC6A5
R-HSA-432030	Transport of glycerol from adipocytes to the liver by Aquaporins	4.7E-02	9.5E-02	AQP7
R-HSA-1226099	Signaling by FGFR in disease	5.4E-02	9.7E-02	KRAS;FGFR3

Table 4.3 Top Common Pathways Enriched with Genes Influencing Treatment Response in all Oral Cancer Patients

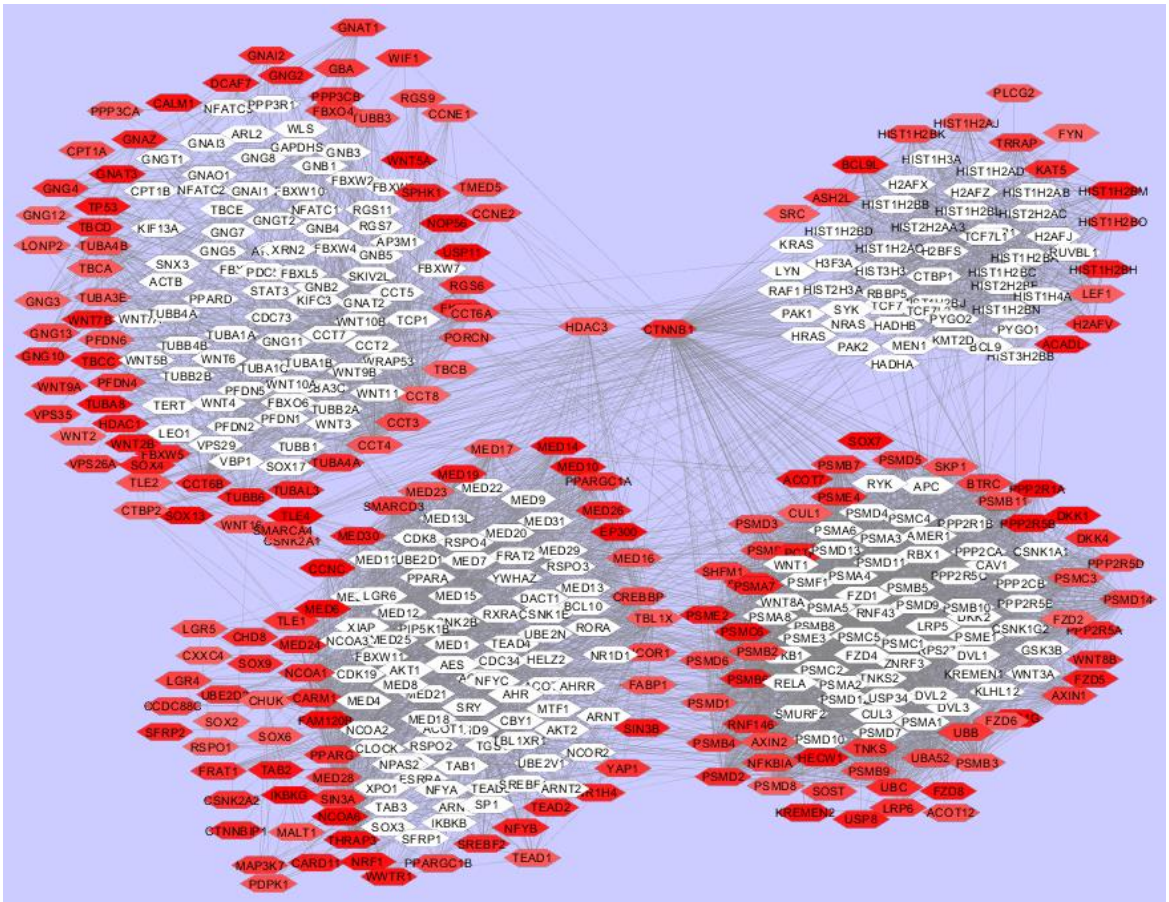
Pathway name	p-value	FDR	Influential Genes Enriched in Pathway
Signaling by WNT	6.7E-15	5.2E-12	HIST1H2BM;HIST1H2BK;CAMK2A;ITPR2;LRP6;PPP3CA; PPP3CB;GNG2;PSMB3; PPP2R1A;PSMD2;PSMB1;PSMD1;SOST;SOX6;BCL9L;SKP1;CSNK2A1; HIST1H2AJ;WNT5A;H2AFV;PPP2R5D;WNT16;RNF146;PSMC3;PSME4
CLEC7A (Dectin-1) signaling	8.5E-08	7.3E-06	PPP3CA;PPP3CB;PSMC3;PSMB3;PSMD2;PSMB1;PSME4;ITPR2; PSMD1;BCL10;MALT1;SKP1
Cooperation of PDCL (PhLP1) and TRiC/CCT in G-protein beta folding	3.3E-05	4.7E-04	GNG2;CSNK2A1;CCT8;RGS6;CCT6B;CCT4
Calcineurin activates NFAT	0.01	0.03	PPP3CA;PPP3CB

Table 4.4. Drug Leads identified in FDA approved and Traditional Chinese Medicine database

Table 4	Drug Leads by Population Subset, Linked Gene, and Tanimoto Similarity Score			
Reference Ligand	RCSB linked Protein/Gene	Drug Candidates FDA (Disease Treated)	TCM Candidates	Tanimoto Score (FDA),(TCM)
(2S)-3-{[{[(2S)-2,3-DIHYDROXYPROPYL]OXY}(HYDROXY)PHOSPHORYL]OXY}-2-[(6E)-HEXADEC-6-ENOYLOXY]PROPYL (8E)-OCTADEC-8-ENOATE	CTNNB1	No candidates found	No candidates found	<50 %
5'-GUANOSINE-DIPHOSPHATE-MONOTHIOPHOSPHATE	PLCG2	Fludarabine(Lung Cancer), Inosine (Multiple Sclerosis), Ganciclovir(HIV), Didanosine(HIV),Entecovir(HepB,HIV)	[4-(2,6-dimethylmorpholin-4-yl)sulfonylphenyl]-[4-(2-phenoxyethyl)piperazin-1-yl]-methanone	(60 %, 59%, 56%, 53%, 51%), (55%)
CYTOSINE ARABINOSE-5'-PHOSPHATE	UBB	Cytarabine(Cancer),Fludarabine(Cancer),Azacitidine(myelodysplastic syndrome),Gemcitabine(Cancer), Lamivudine(HIV)	No Candidates Found	(80 %, 73%, 72%, 69%, 63%),(NA)

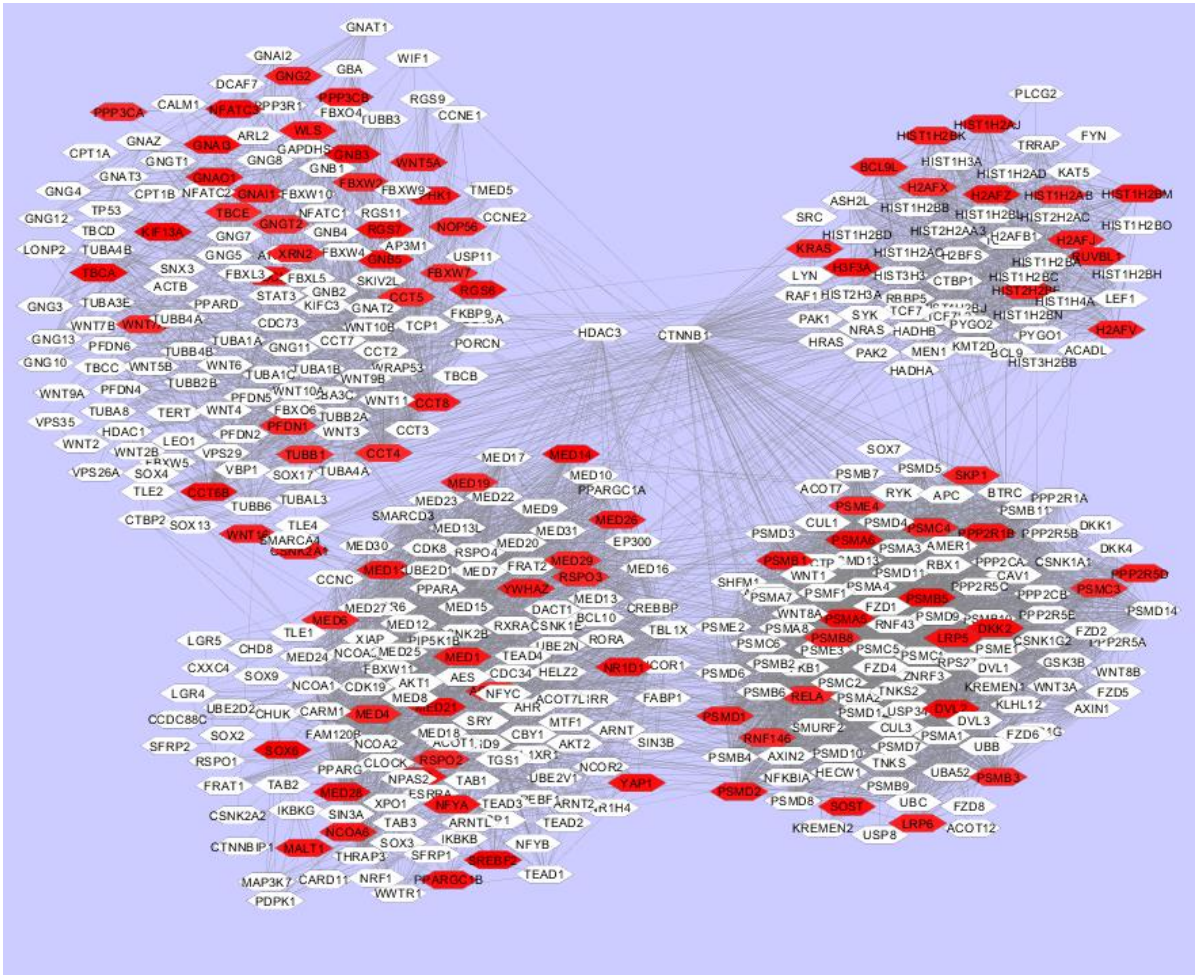
d-MYO-INOSITOL-1,4,5,6-TETRAKISPHOSPHATE	HDAC3	No Candidates	Mallotophillipen-D, Quercetin, Diosmetin	(NA), (77%, 77%, 77%)

Figure 4.1. Network Visualization of Pathways Enriched with Genes Influencing Platinum Based Treatment Response in Oral Cancer



Notes: Hexagon shapes are genes. Dark red are of greater influence based upon random forest analysis results (Within top 5% of influential genes), white genes do not fall within the criteria of being in the to 5% of influential genes. Genes are clustered by pathway with Calcineurin pathway in the top left corner, The WNT signaling pathway in the upper right corner, CLEC7A pathway genes are in the lower right corner, and the cooperation of PDCL (PhLP1) and TRiC/CCT in G-protein beta folding pathway is in the lower left corner. Grey lines represent interconnectedness between pathways. Genes are arranged in a flower pattern with influential genes on the outside and genes that are not influential in treatment response on the inside. Patterns are constructed to contrast gene influence with those patients not receiving platinum based chemotherapy.

Figure 4.2. Network Visualization of Pathways Enriched with Genes Influencing Non-Platinum Based Treatment Response in Oral Cancer



Notes: Hexagon shapes are genes. Dark red are of greater influence based upon random forest analysis results (Within top 5% of influential genes), white genes do not fall within the criteria of being in the to 5% of influential genes. Loss of ring structure is indicative of differences in influence of genes between patients receiving platinum and non-platinum therapies. Genes are clustered by pathway with Calcineurin pathway in the top left corner, The WNT signaling pathway in the upper right corner, CLEC7A pathway genes are in the lower right corner, and the cooperation of PDCL (PhLP1) and TRiC/CCT in G-protein beta folding pathway is in the lower left corner. Grey lines represent interconnectedness between pathways.

Chapter 5

Conclusion

The studies detailed in depth in this dissertation contribute to the current body of knowledge on the determinants of health in oral cancer. There is great variability in the response to treatment in patients with oral cancer. Agencies assigning quality of care guidelines such as NCCN provide recommendations for when providers should use chemotherapy in oral cancer patients. This study provides evidence that there is merit to the recommendations of NCCN in situations when patients meet criteria for chemotherapy, and when they do not. The use of a large population based sample of ethnically diverse patients show that NCCN guideline adherence is an important predictor of patient survival. Subset analysis provides evidence that assigning chemotherapy treatment to those patients with large tumors that receive radiation only (possibly indicating that such tumors are inoperable) were significantly less likely to survive when prescribed chemotherapy in contradiction of NCCN recommendations. These results provide guidance to the NCCN to promote stronger language against the use of chemotherapy in node negative patients of stage three and lower. Most importantly this work has validated the importance of adherence to NCCN guidelines on the use of Chemotherapy for oral cancer patients.

In addition to the validation of NCCN guidelines, this dissertation has specifically focused on the use of gene expression signatures predicting response to treatment in oral cancer patients. In contrast to the initial study reliant upon the well powered California Cancer Registry, the study of oral cancer patients from the Cancer genome atlas was small with only 276 cases used to detect those genes differentially expressed over 100 runs of differential expression analysis. By creating an aggregated gene signature, generated from the top differentially expressed genes a gene set that predicted treatment response in oral cancer was identified. This model strengthens evidence that the ligand channel gating pathways are influential in the classification of treatment responders. An example of how small datasets representing rare cancers is also set, highlighting the utility of Monte Carlo cross validation is such situations. While the availability of public genetic data is growing, this growth is naturally limited in cancers where incidence is lower. Thus, validating a method that successfully identifies gene expression signatures composed of the differentially expressed genes provides value to other researchers seeking to validate gene signatures in rare cancers where the number of observations are sparse. By utilizing the high dimensional data in the cancer genome atlas, signature performance was described across gender, stage, and grade visualizing the effectiveness of the signature across meaningful strata. This analysis describes the genes and pathways that are significant predictors of treatment response in oral cancer, with strongest evidence supporting the role of ligand gated ion channel pathways.

The final chapter of this dissertation approaches the problem of interpreting high dimensional data with low number of observations by utilizing non parametric machine learning method known as random forest. This method of classification can be utilized to rank genes in their ability to classify treatment responders and non-responders. This study shows those genes and pathways that are influential in predicting treatment response in oral cancer. Specifically, it shows those unique and shared pathways between those patients receiving platinum based chemotherapy and non-platinum based chemotherapy. It also shows that the pathways that are influential in both treatment subgroups (platinum and non-platinum) contain influential genes that differ depending on treatment. This means that the genes that influence treatment response in the WNT signaling pathway, or in pathways regulating histone deacetylation pathway that were based upon platinum based chemotherapy patient gene expression data are not all the same genes in the ketone body metabolism pathway that influence treatment response in non-platinum based chemotherapy. This method allows the targeting of unique drivers of treatment response that may only be relevant to patients on a particular treatment.

In addition to identifying influential gene pathways this study had integrated the use of a chemical informatics package in R to also identify drug leads that may act on such targets. This process aired with a review of the pathway for highly connected genes identified in Cytoscape allowed for the selection of small molecules with high Tanimoto similarity when compared to reference molecules that are known to bind to proteins expressed by the gene found to be influential via integrated analyses. When pairing

these drug leads with existing literature it was found that this approach does identify drug leads that are effective in reducing progression of oral cancer cells or inducing apoptosis when tested in cell lines. This provides validation of this method for the identification of drug leads in rare cancers with relatively low number of observations.

In total, this dissertation has used classical epidemiologic analyses to identify the benefit of adherence to NCCN guidelines, adopted permutation of RNA seq data to strengthen the signal of an oral cancer gene signature, and integrated random forest analysis with gene network and chemical informatics analysis to identify influential genes/pathways and possible drug leads that could act on such genes. This work advances the knowledge of effective treatments in oral cancer, and provides greater understanding of the genetic pathways influencing treatment response. The methods outlined in this study can be adopted by others when faced with the challenge of extracting meaning from high dimensional datasets with small number of observations. The findings related to adherence to NCCN guidelines are relevant specifically to providers treating patients with oral cancer. This study also serves as an impetus for further research investigating the action of specific small molecules on oral cancer treatment response. All Chapters within this manuscript have focused upon the central theme of the identification of the determinants of health in patients with oral cancer. The spirit embodied within this manuscript is one of positive change, positive change in advancing the standards of care in oral cancer, change in rigor to the approach applied to identifying those genes influencing rare cancers, and change to the pace at which the

identification of novel treatments for this disease must be identified not for the cancer patients of tomorrow but for those being treated in the here and now.

Bibliography

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin 2016;66:7–30. doi:10.3322/caac.21332.

[2] Siegel R, Miller K, Jemal A. Cancer statistics , 2015 . CA Cancer J Clin 2015;65:29. doi:10.3322/caac.21254.

[3] Cooper JS, Porter K, Mallin K, Hoffman HT, Weber RS, Ang KK, et al. National cancer database report on cancer of the head and neck: 10-Year update. Head Neck 2009;31:748–58. doi:10.1002/hed.21022.

[4] Sinha P, Logan HL, Mendenhall WM. Human papillomavirus, smoking, and head and neck cancer. Am J Otolaryngol 2012;33:130–6. doi:10.1016/j.amjoto.2011.02.001.

[5] Sturgis EM, Cinciripini PM. Trends in head and neck cancer incidence in relation to smoking prevalence: An emerging epidemic of human papillomavirus-associated cancers? Cancer 2007;110:1429–35. doi:10.1002/cncr.22963.

[6] Ryerson AB, Eheman CR, Altekruse SF, Ward JW, Jemal A, Sherman RL, et al. Annual Report to the Nation on the Status of Cancer, 1975-2012, featuring the increasing incidence of liver cancer. Cancer 2016;122:1312–37.

doi:10.1002/cncr.29936.

[7] Ragin CCR, Taioli E. Survival of squamous cell carcinoma of the head and neck in relation to human papillomavirus infection: Review and meta-analysis. Int J Cancer 2007;121:1813–20. doi:10.1002/ijc.22851.

[8] Bosch FX, Lorincz A, Muñoz N, Meijer CJLM, Shah K V. The causal relation between human papillomavirus and cervical cancer. J Clin Pathol 2002;55:244–65. doi:10.1136/jcp.55.4.244.

[9] Daley E, Dodd V, DeBate R, Vamos C, Wheldon C, Kline N, et al. Prevention of HPV-related oral cancer: Assessing dentists' readiness. Public Health 2014;128:231–8. doi:10.1016/j.puhe.2013.12.002.

[10] D'Souza G, Gross ND, Pai SI, Haddad R, Anderson KS, Rajan S, et al. Oral human papillomavirus (HPV) infection in HPV-positive patients with oropharyngeal cancer and their partners. J Clin Oncol 2014;32:2408–15. doi:10.1200/JCO.2014.55.1341.

[11] Lee SY, Cho NH, Choi EC, Baek SJ, Kim WS, Shin DH, et al. Relevance of human papilloma virus (HPV) infection to carcinogenesis of oral tongue cancer. Int J Oral Maxillofac Surg 2010;39:678–83. doi:10.1016/j.ijom.2010.03.014.

[12] Marur S, D'Souza G, Westra WH, Forastiere AA. HPV-associated head and neck cancer: A virus-related cancer epidemic. Lancet Oncol 2010;11:781–9. doi:10.1016/S1470-2045(10)70017-6.

[13] Radosevich JA. HPV and cancer. vol. 9789400754. 2012. doi:10.1007/978-94-007-5437-9.

[14] Wood O, Woo J, Seumois G, Savelyeva N, McCann KJ, Singh D, et al. Gene

expression analysis of TIL rich HPV-driven head and neck tumors reveals a distinct B-cell signature when compared to HPV independent tumors. Oncotarget 2016;5:10788. doi:10.18632/oncotarget.10788.

[15] Keck MK, Zuo Z, Khattri A, Stricker TP, Brown CD, Imanguli M, et al. Integrative analysis of head and neck cancer identifies two biologically distinct HPV and three non-HPV subtypes. Clin Cancer Res 2015;21:870–81. doi:10.1158/1078-0432.CCR-14-2481.

[16] Chakravarthy A, Henderson S, Thirdborough SM, Ottensmeier CH, Su X, Lechner M, et al. Human papillomavirus drives tumor development throughout the head and neck: Improved prognosis is associated with an immune response largely restricted to the Oropharynx. J Clin Oncol 2016;34:4132–41. doi:10.1200/JCO.2016.68.2955.

[17] Hayes DN, Van Waes C, Seiwert TY. Genetic landscape of human papillomavirus-associated head and neck cancer and comparison to tobacco-related tumors. J Clin Oncol 2015;33:3227–34. doi:10.1200/JCO.2015.62.1086.

[18] Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, et al. The mutational landscape of head and neck squamous cell carcinoma. Science 2011;333:1157–60. doi:10.1126/science.1208130.

[19] Lassen P, Eriksen JG, Hamilton-Dutoit S, Tramm T, Alsner J, Overgaard J. HPV-associated p16-expression and response to hypoxic modification of radiotherapy in head and neck cancer. Radiother Oncol 2010;94:30–5. doi:10.1016/j.radonc.2009.10.008.

[20] Lassen P, Eriksen JG, Hamilton-Dutoit S, Tramm T, Alsner J, Overgaard J. Effect

of HPV-associated p16INK4A expression on response to radiotherapy and survival in squamous cell carcinoma of the head and neck. J Clin Oncol 2009;27:1992–8. doi:10.1200/JCO.2008.20.2853.

[21] Lassen P, Eriksen JG, Krogdahl A, Therkildsen MH, Ulhøi BP, Overgaard M, et al. The influence of HPV-associated p16-expression on accelerated fractionated radiotherapy in head and neck cancer: Evaluation of the randomised DAHANCA 6&7 trial. Radiother Oncol 2011;100:49–55. doi:10.1016/j.radonc.2011.02.010.

[22] Hashibe M, Brennan P, Chuang S, Boccia S, Castellsague X, Chen C, et al. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the INHANCE consortium. Cancer Epidemiol Biomarkers Prev 2009;18:541–50. doi:10.1158/1055-9965.EPI-08-0347.

[23] Hecht SS. Tobacco smoke carcinogens and lung cancer. Curr Cancer Res 2011;6:53–74. doi:10.1007/978-1-61737-995-6_3.

[24] Hecht SS. Tobacco carcinogens, their biomarkers and tobacco-induced cancer. Nat Rev Cancer 2003;3:733–44. doi:10.1038/nrc1190.

[25] Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene 2002;21:7435–51. doi:10.1038/sj.onc.1205803.

[26] Hecht SS. Cigarette smoking: Cancer risks, carcinogens, and mechanisms. Langenbeck's Arch Surg 2006;391:603–13. doi:10.1007/s00423-006-0111-z.

[27] Amos a., Greaves L, Nichter M, Bloch M. Women and tobacco: a call for including gender in tobacco control research, policy and practice. Tob Control 2012;21:236–43. doi:10.1136/tobaccocontrol-2011-050280.

[28] Weatherspoon DJ, Chattopadhyay A, Boroumand S, Garcia I. Oral cavity and oropharyngeal cancer incidence trends and disparities in the United States: 2000-2010. Cancer Epidemiol 2015;39:497–504. doi:10.1016/j.canep.2015.04.007.

[29] Hashibe M, Brennan P, Chuang S-C, Boccia S, Castellsague X, Chen C, et al. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. Cancer Epidemiol Biomarkers Prev 2009;18:541–50. doi:10.1158/1055-9965.EPI-08-0347.

[30] Vidal L, Gillison ML. Human Papillomavirus in HNSCC: Recognition of a Distinct Disease Type. Hematol Oncol Clin North Am 2008;22:1125–42. doi:10.1016/j.hoc.2008.08.006.

[31] Akervall J, Nandalur S, Zhang J, Qian C-N, Goldstein N, Gyllerup P, et al. A novel panel of biomarkers predicts radioresistance in patients with squamous cell carcinoma of the head and neck. Eur J Cancer 2014;50:570–81. doi:10.1016/j.ejca.2013.11.007.

[32] Pramana J, Van Den Brekel MWM, Van Velthuysen MLF, Wessels LFA, Nuyten DS, Hofland I, et al. Gene expression profiling to predict outcome after chemoradiation in head an neck cancer. Int J Radiat Oncol Biol Phys 2007;69:1544–52.

[33] Dumur CI, Ladd AC, Wright H V, Penberthy LT, Wilkinson DS, Powers CN, et al. Genes involved in radiation therapy response in head and neck cancers. Laryngoscope 2009;119:91–101. doi:10.1002/lary.20005.

[34] Wong N, Khwaja SS, Baker CM, Gay HA, Thorstad WL, Daly MD, et al.

Prognostic microRNA signatures derived from The Cancer Genome Atlas for head and neck squamous cell carcinomas. Cancer Med 2016:1–10. doi:10.1002/cam4.718.

[35] Kimple RJ, Smith MA, Blitzer GC, Torres AD, Martin JA, Yang RZ, et al. Enhanced radiation sensitivity in HPV-positive head and neck cancer. Cancer Res 2013;73:4791–800. doi:10.1158/0008-5472.CAN-13-0587.

[36] Dayyani F, Etzel CJ, Liu M, Ho C-H, Lippman SM, Tsao AS. Meta-analysis of the impact of human papillomavirus (HPV) on cancer risk and overall survival in head and neck squamous cell carcinomas (HNSCC). Head Neck Oncol 2010;2:15. doi:10.1186/1758-3284-2-15.

[37] Sørensen BS, Busk M, Olthof N, Speel EJ, Horsman MR, Alsner J, et al. Radiosensitivity and effect of hypoxia in HPV positive head and neck cancer cells. Radiother Oncol 2013;108:500–5. doi:10.1016/j.radonc.2013.06.011.

[38] Nagel R, Martens-De Kemp SR, Buijze M, Jacobs G, Braakhuis BJM, Brakenhoff RH. Treatment response of HPV-positive and HPV-negative head and neck squamous cell carcinoma cell lines. Oral Oncol 2013;49:560–6. doi:10.1016/j.oraloncology.2013.03.446.

[39] Park SM, Lim MK, Shin SA, Yun YH. Impact of prediagnosis smoking, alcohol, obesity, and insulin resistance on survival in male cancer patients: National Health Insurance Corporation Study. J Clin Oncol 2006;24:5017–24. doi:10.1200/JCO.2006.07.0243.

[40] Daraei P, Moore CE. Racial Disparity Among the Head and Neck Cancer Population. J Cancer Educ 2015;30:546–51. doi:10.1007/s13187-014-0753-4.

[41] Moore CE, Warren R, Maclin SD. Head and neck cancer disparity in underserved communities: probable causes and the ethics involved. J Health Care Poor Underserved 2012;23:88–103. doi:10.1353/hpu.2012.0165.

[42] P. D. Racial Disparity Among the Head and Neck Cancer Population. J Cancer Educ 2015;30:546–51.

[43] Schrank TP, Han Y, Weiss H, Resto VA. Case-matching analysis of head and neck squamous cell carcinoma in racial and ethnic minorities in the United States--possible role for human papillomavirus in survival disparities. Head Neck 2011;33:45–53. doi:10.1002/hed.21398.

[44] Daraei P, Moore CE, P. D. Racial Disparity Among the Head and Neck Cancer Population. J Cancer Educ 2015;30:546–51.

[45] Ragin CC, Langevin SM, Marzouk M, Grandis J, Taioli E. Determinants of head and neck cancer survival by race. Head Neck 2011;33:1092–8. doi:10.1002/hed.21584.

[46] National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Head and Neck. Version 12016 2016:1–174.

[47] Daly B, Olopade OI. A perfect storm: How tumor biology, genomics, and health care delivery patterns collide to create a racial survival disparity in breast cancer and proposed interventions for change. CA Cancer J Clin 2015;65:221–38. doi:10.3322/caac.21271.

[48] Simpson DR, Martínez ME, Gupta S, Hattangadi-Gluth J, Mell LK, Heestand G, et al. Racial disparity in consultation, treatment, and the impact on survival in metastatic colorectal cancer. J Natl Cancer Inst 2013;105:1814–20.

doi:10.1093/jnci/djt318.

[49] Griggs JJ, Sorbero MES, Stark AT, Heininger SE, Dick AW. Racial disparity in the dose and dose intensity of breast cancer adjuvant chemotherapy. Breast Cancer Res Treat 2003;81:21–31. doi:10.1023/A:1025481505537.

[50] Underwood W, Dunn RL, Williams C, Lee CT. Gender and geographic influence on the racial disparity in bladder cancer mortality in the US. J Am Coll Surg 2006;202:284–90. doi:10.1016/j.jamcollsurg.2005.09.009.

[51] Howlader N, Noone AM, Krapcho M, Neyman N, Aminou R, Waldron W, et al. SEER Cancer Statistics Review, 1975–2010, National Cancer Institute. Bethesda, MD, based on November 2012 SEER data submission, posted to the SEER web site, 2013. http//seer.cancer.gov/csr/1975_2010 (Accessed June 08, 2013) 2013.

[52] Weng Y, Korte JE. Racial disparities in being recommended to surgery for oral and oropharyngeal cancer in the United States. Community Dent Oral Epidemiol 2012;40:80–8. doi:10.1111/j.1600-0528.2011.00638.x.

[53] Shiboski CH, Schmidt BL, Jordan RCK. Racial disparity in stage at diagnosis and survival among adults with oral cancer in the US. Community Dent Oral Epidemiol 2007;35:233–40. doi:10.1111/j.0301-5661.2007.00334.x.

[54] Settle K, Posner MR, Schumaker LM, Tan M, Suntharalingam M, Goloubeva O, et al. Racial survival disparity in head and neck cancer results from low prevalence of human papillomavirus infection in black oropharyngeal cancer patients. Cancer Prev Res 2009;2:776–81. doi:10.1158/1940-6207.CAPR-09-0149.

[55] Nesbitt S, Victor RG. Pathogenesis of hypertension in African Americans. Congest Heart Fail 2004;10:24–9. doi:10.1111/j.1527-5299.2004.02021.x.

[56] Marshall MC. Diabetes in African Americans. Postgrad Med J 2005;81:734–40. doi:10.1136/pgmj.2004.028274.

[57] Diaz KM, Veerabhadrappa P, Brown MD, Whited MC, Dubbert PM, Hickson D a. Prevalence, Determinants, and Clinical Significance of Masked Hypertension in a Population-Based Sample of African Americans: The Jackson Heart Study. Am J Hypertens 2014;28:1–9. doi:10.1093/ajh/hpu241.

[58] Ayanian JZ. Determinants of racial and ethnic disparities in surgical care. World J Surg 2008;32:509–15. doi:10.1007/s00268-007-9344-4.

[59] Britton B V., Nagarajan N, Zogg CK, Selvarajah S, Schupper AJ, Kironji AG, et al. Awareness of racial/ethnic disparities in surgical outcomes and care: factors affecting acknowledgment and action. Am J Surg 2016;212:102–8.e2. doi:10.1016/j.amjsurg.2015.07.022.

[60] Silber JH, Rosenbaum PR, Kelz RR, Gaskin DJ, Ludwig JM, Ross RN, et al. Examining causes of racial disparities in general surgical mortality. Med Care 2015;53:619–29. doi:10.1097/MLR.0000000000000377.

[61] Bristow R, Chang J, Ziogas A, Anton-Culver H. NCCN treatment guidelines for ovarian cancer: A population-based vali-dation study of structural and process quality measures. Gynecol Oncol 2013;130:e18. doi:10.1016/j.ygyno.2013.04.104.

[62] Yamamoto N, Shibahara T. Epidemiology of the oral cancer. Oral Cancer Diagnosis Ther., 2015, p. 1–21. doi:10.1007/978-4-431-54938-3_1.

[63] Méndez E, Houck JR, Doody DR, Fan W, Lohavanichbutr P, Rue TC, et al. A genetic expression profile associated with oral cancer identifies a group of

patients at high risk of poor survival. Clin Cancer Res 2009;15:1353–61. doi:10.1158/1078-0432.CCR-08-1816.

[64] Saintigny P, Zhang L, Fan Y-H, El-Naggar AK, Papadimitrakopoulou V, Feng L, et al. Gene Expression Profiling Predicts the Development of Oral Cancer. Cancer Prev Res (Phila) 2011;4:218–29. doi:10.1158/1940-6207.CAPR-10-0155.

[65] Sakamoto K, Aragaki T, Morita K ichi, Kawachi H, Kayamori K, Nakanishi S, et al. Down-regulation of keratin 4 and keratin 13 expression in oral squamous cell carcinoma and epithelial dysplasia: A clue for histopathogenesis. Histopathology 2011;58:531–42. doi:10.1111/j.1365-2559.2011.03759.x.

[66] Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Wspolczesna Onkol 2015;1A:A68–77. doi:10.5114/wo.2014.47136.

[67] Wan YW, Mach CM, Allen GI, Anderson ML, Liu Z. On the reproducibility of TCGA ovarian cancer microRNA profiles. PLoS One 2014;9. doi:10.1371/journal.pone.0087782.

[68] Zhang W. TCGA divides gastric cancer into four molecular subtypes: Implications for individualized therapeutics. Chin J Cancer 2014;33:469–70. doi:10.5732/cjc.014.10117.

[69] Bloom S. TCGA Analysis Reveals New Insights about Colorectal Cancer. 2012.

[70] Rios Velazquez E, Meier R, Dunn Jr WD, Alexander B, Wiest R, Bauer S, et al. Fully automatic GBM segmentation in the TCGA-GBM dataset: Prognosis and correlation with VASARI features. Sci Rep 2015;5:16822. doi:10.1038/srep16822.

[71] News. TCGA Sees Heterogeneity in Head and Neck Cancers. Cancer Discov

2013;3:475–6. doi:10.1158/2159-8290.CD-NB2013-049.

[72] Taylor JMG, Ankerst DP, Andridge RR. Validation of biomarker-based risk prediction models. Clin Cancer Res 2008;14:5977–83. doi:10.1158/1078-0432.CCR-07-4534.

[73] Zhang P. Model Selection Via Multifold Cross Validation. Ann Stat 1993;21:299–313. doi:doi:10.1214/aos/1176349027.

[74] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) (9780387848570): Trevor Hastie, Robert Tibshirani, Jerome Friedman: Books. Elem. Stat. Learn. dta mining, inference, Predict., 2011, p. 501–20.

[75] Mishra D, Sahu B. Feature selection for cancer classification: a signal-to-noise ratio approach. Int J Sci Eng Res 2011;2:1–7.

[76] Stephenson AJ, Smith A, Kattan MW, Satagopan J, Reuter VE, Scardino PT, et al. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. Cancer 2005;104:290–8. doi:10.1002/cncr.21157 [doi].

[77] Park S, Shimizu C, Shimoyama T, Takeda M, Ando M, Kohno T, et al. Gene expression profiling of ATP-binding cassette (ABC) transporters as a predictor of the pathologic response to neoadjuvant chemotherapy in breast cancer patients. Breast Cancer Res Treat 2006;99:9–17. doi:10.1007/s10549-006-9175-2.

[78] Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. Tissue classification with gene expression profiles. J Comput Biol 2000;7:559–83. doi:10.1089/106652700750050943.

[79] Zhu C-Q, Strumpf D, Li C-Y, Li Q, Liu N, Der S, et al. Prognostic Gene Expression Signature for Squamous Cell Carcinoma of Lung. Clin Cancer Res 2010;16:5038–48. doi:10.1158/1078-0432.CCR-10-0612.

[80] Barrier A, Boelle P-Y, Roser F, Gregg J, Tse C, Brault D, et al. Stage II colon cancer prognosis prediction by tumor gene expression profiling. J Clin Oncol 2006;24:4685–91. doi:10.1200/JCO.2005.05.0229.

[81] Patnaik SK, Kannisto E, Knudsen S, Yendamuri S. Evaluation of microRNA expression profiles that may predict recurrence of localized stage I non-small cell lung cancer after surgical resection. Cancer Res 2010;70:36–45. doi:10.1158/0008-5472.CAN-09-3153.

[82] Xu QS, Liang YZ, Du YP. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. J Chemom 2004;18:112–20. doi:10.1002/cem.858.

[83] Collins FS. The Cancer Genome Atlas (TCGA). Online 2007:1–17.

[84] Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol 2013;14:R95. doi:10.1186/gb-2013-14-9-r95.

[85] Shao J. Linear Model Selection by Cross-Validation. J Amer Stat Assoc 1993;88:486–94.

[86] Li T, Tang W, Zhang L. Monte Carlo cross-validation analysis screens pathway cross-talk associated with Parkinson???s disease. Neurol Sci 2016;37:1327–33. doi:10.1007/s10072-016-2595-9.

[87] Colaprico A, Cava C, Bertoli G, Bontempi G, Castiglioni I. Integrative Analysis

with Monte Carlo Cross-Validation Reveals miRNAs Regulating Pathways Cross-Talk in Aggressive Breast Cancer. Biomed Res Int 2015;2015. doi:10.1155/2015/831314.

[88] Robin AX, Turck N, Hainard A, Lisacek F, Sanchez J, Müller M, et al. Package " pROC ." 2012-09-10 09:34:56 2013:1–71. doi:10.1186/1471-2105-12-77.

[89] Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2016;44:D481–7. doi:10.1093/nar/gkv1351.

[90] Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: A knowledgebase of biological pathways. Nucleic Acids Res 2005;33. doi:10.1093/nar/gki072.

[91] Gharahkhani P, Fitzgerald RC, Vaughan TL, Palles C, Gockel I, Tomlinson I, et al. Genome-wide association studies in oesophageal adenocarcinoma and Barrett's oesophagus: a large-scale meta-analysis. Lancet Oncol 2016;17:1363–73. doi:10.1016/S1470-2045(16)30240-6.

[92] Chang KW, Yuan TC, Fang KP, Yang FS, Liu CJ, Chang CS, et al. The increase of voltage-gated potassium channel Kv3.4 mRNA expression in oral squamous cell carcinoma. J Oral PatholMed 2003;32:606–11.

[93] Lew T-S, Chang C-S, Fang K-P, Chen C-Y, Chen C-H, Lin S-C. The involvement of Kv3.4 voltage-gated potassium channel in the growth of an oral squamous cell carcinoma cell line\rdoi:10.1111/j.1600-0714.2004.00236.x. J Oral Pathol Med 2004;33:543–9.

[94] Fernández-Valle Á, Rodrigo JP, García-Pedrero JM, Rodríguez-Santamarta T,

Allonca E, Lequerica-Fernández P, et al. Expression of the voltage-gated potassium channel Kv3.4 in oral leucoplakias and oral squamous cell carcinomas. Histopathology 2016;69:91–8. doi:10.1111/his.12917.

[95] Bourguignon LYW, Wong G, Earle C, Chen L. Hyaluronan-CD44v3 interaction with Oct4-Sox2-Nanog promotes miR-302 expression leading to self-renewal, clonal formation, and cisplatin resistance in cancer stem cells from head and neck squamous cell carcinoma. J Biol Chem 2012;287:32800–24. doi:10.1074/jbc.M111.308528.

[96] Huang CE, Yu CC, Hu FW, Chou MY, Tsai LL. Enhanced Chemosensitivity by Targeting Nanog in Head and Neck Squamous Cell Carcinomas. Int J Mol Sci 2014;15:14935–48. doi:10.3390/ijms150914935.

[97] Schaaij-Visser TBM, Bremmer JF, Braakhuis BJM, Heck AJR, Slijper M, van der Waal I, et al. Evaluation of cornulin, keratin 4, keratin 13 expression and grade of dysplasia for predicting malignant progression of oral leukoplakia. Oral Oncol 2010;46:123–7. doi:10.1016/j.oraloncology.2009.11.012.

[98] Hamakawa H, Fukuzumi M, Bao Y, Sumida T, Kayahara H, Onishi A, et al. Keratin mRNA for detecting micrometastasis in cervical lymph nodes of oral cancer. Cancer Lett 2000;160:115–23. doi:10.1016/S0304-3835(00)00574-7.

[99] Becker KG, Hosack DA, Dennis G, Lempicki RA, Bright TJ, Cheadle C, et al. PubMatrix: a tool for multiplex literature mining. BMC Bioinformatics 2003;4:61. doi:10.1186/1471-2105-4-61.

[100] TCGA Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature 2015;517:576–82. doi:10.1038/nature14129.

[101] Breiman L (University of C. Random forest. Mach Learn 1999;45:1–35. doi:10.1023/A:1010933404324.

[102] Qi Y. Random Forest for Bioinformatics. Ensemble Mach Learn 2012:307–23. doi:10.1007/978-1-4419-9326-7_11.

[103] Breiman L. Random forests. Mach Learn 2001;45:5–32. doi:10.1023/A:1010933404324.

[104] Chen X, Wang L, Ishwaran H. An integrative pathway-based clinical-genomic model for cancer survival prediction. Stat Probab Lett 2010;80:1313–9. doi:10.1016/j.spl.2010.04.011.

[105] Mutanga O, Adam E, Cho MA. High density biomass estimation for wetland vegetation using worldview-2 imagery and random forest regression algorithm. Int J Appl Earth Obs Geoinf 2012;18:399–406. doi:10.1016/j.jag.2012.03.012.

[106] Tong W, Xie Q, Hong H, Fang H, Shi L, Perkins R, et al. Using decision forest to classify prostate cancer samples on the basis of SELDI-TOF MS data: Assessing chance correlation and prediction confidence. Environ Health Perspect 2004;112:1622–7. doi:10.1289/txg.7109.

[107] Chen XW, Liu M. Prediction of protein-protein interactions using random decision forest framework. Bioinformatics 2005;21:4394–400. doi:10.1093/bioinformatics/bti721.

[108] Phillips M, Cataneo RN, Ditkoff BA, Fisher P, Greenberg J, Gunawardena R, et al. Prediction of breast cancer using volatile biomarkers in the breath. Breast Cancer Res Treat 2006;99:19–21. doi:10.1007/s10549-006-9176-1.

[109] Pang H, Zhao H. Stratified pathway analysis to identify gene sets associated with

oral contraceptive use and breast cancer. Cancer Inform 2014;13:73–8. doi:10.4137/CIN.S13973.

[110] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008;2:841–60. doi:10.1214/08-AOAS169.

[111] Williams A, Tkachenko V. The Royal Society of Chemistry and the delivery of chemistry data repositories for the community. J Comput Aided Mol Des 2014;28:1023–30. doi:10.1007/s10822-014-9784-5.

[112] Cao DS, Xiao N, Xu QS, Chen AF. Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. Bioinformatics 2015;31:279–81. doi:10.1093/bioinformatics/btu624.

[113] Monteith GR, McAndrew D, Faddy HM, Roberts-Thomson SJ. Calcium and cancer: targeting Ca2+ transport. Nat Rev Cancer 2007;7:519–30. doi:10.1038/nrc2171.

[114] Pahor M, Guralnik JM, Salive ME, Corti MC, Carbonin P, Havlik RJ. Do calcium channel blockers increase the risk of cancer? Am J Hypertens 1996;9:695–9. doi:10.1016/0895-7061(96)00186-0.

[115] Pahor M, Guralnik JM, Ferrucci L, Corti MC, Salive ME, Cerhan JR, et al. Calcium-channel blockade and incidence of cancer in aged populations. Lancet 1996;348:493–7. doi:10.1016/S0140-6736(96)04277-8.

[116] Giudice FS, Pinto DS, Nör JE, Squarize CH, Castilho RM. Inhibition of Histone Deacetylase Impacts Cancer Stem Cells and Induces Epithelial-Mesenchyme Transition of Head and Neck Cancer. PLoS One 2013;8. doi:10.1371/journal.pone.0058672.

[117] Erlich RB, Kherrouche Z, Rickwood D, Endo-Munoz L, Cameron S, Dahler A, et al. Preclinical evaluation of dual PI3K-mTOR inhibitors and histone deacetylase inhibitors in head and neck squamous cell carcinoma. Br J Cancer 2012;106:107–15. doi:10.1038/bjc.2011.495.

[118] Haigentz M, Kim M, Sarta C, Lin J, Keresztes RS, Culliney B, et al. Phase II trial of the histone deacetylase inhibitor romidepsin in patients with recurrent/metastatic head and neck cancer. Oral Oncol 2012;48:1281–8. doi:10.1016/j.oraloncology.2012.05.024.

[119] Taipale J, Beachy PA. The Hedgehog and Wnt signalling pathways in cancer. Nature 2001;411:349–54. doi:10.1038/35077219.

[120] Gulino A, Ferretti E, De Smaele E. Hedgehog signalling in colon cancer and stem cells. EMBO Mol Med 2009;1:300–2. doi:10.1002/emmm.200900042.

[121] Watkins DN, Berman DM, Burkholder SG, Wang B, Beachy PA, Baylin SB. Hedgehog signalling within airway epithelial progenitors and in small-cell lung cancer. Nature 2003;422:313–7. doi:10.1038/nature01493.

[122] Shaw G, Price AM, Ktori E, Bisson I, Purkis PE, McFaul S, et al. Hedgehog Signalling in Androgen Independent Prostate Cancer. Eur Urol 2008;54:1333–43. doi:10.1016/j.eururo.2008.01.070.

[123] Kasper M, Jaks V, Fiaschi M, Toftgård R. Hedgehog signalling in breast cancer. Carcinogenesis 2009;30:903–11. doi:10.1093/carcin/bgp048.

[124] Thayer SP, Di Magliano MP, Heiser PW, Nielsen CM, Roberts DJ, Lauwers GY, et al. Hedgehog is an early and late mediator of pancreatic cancer tumorigenesis. Nature 2003;425:851–6. doi:10.1038/nature02009.

[125] Parikh N, Hilsenbeck S, Creighton CJ, Dayaram T, Shuck R, Shinbrot E, et al. Effects of TP53 mutational status on gene expression patterns across 10 human cancer types. J Pathol 2014;232:522–33. doi:10.1002/path.4321.

[126] Touat M, Ileana E, Postel-Vinay S, André F, Soria JC. Targeting FGFR signaling in cancer. Clin Cancer Res 2015;21:2684–94. doi:10.1158/1078-0432.CCR-14-2329.

[127] Wynes MW, Hinz TK, Gao D, Martini M, Marek LA, Ware KE, et al. FGFR1 mRNA and protein expression, not gene copy number, predict FGFR TKI sensitivity across all lung cancer histologies. Clin Cancer Res 2014;20:3299–309. doi:10.1158/1078-0432.CCR-13-3060.

[128] Cerliani JP, Vanzulli SI, Piñero CP, Bottino MC, Sahores A, Nuñez M, et al. Associated expressions of FGFR-2 and FGFR-3: From mouse mammary gland physiology to human breast cancer. Breast Cancer Res Treat 2012;133:997–1008. doi:10.1007/s10549-011-1883-6.

[129] Huang X, Jan LY. Targeting potassium channels in cancer. J Cell Biol 2014;206:151–62. doi:10.1083/jcb.201404136.

[130] Bonnet S, Archer SL, Allalunis-Turner J, Haromy A, Beaulieu C, Thompson R, et al. A Mitochondria-K+ Channel Axis Is Suppressed in Cancer and Its Normalization Promotes Apoptosis and Inhibits Cancer Growth. Cancer Cell 2007;11:37–51. doi:10.1016/j.ccr.2006.10.020.

[131] Chang K-W, Yuan T-C, Fang K-P, Yang F-S, Liu C-J, Chang C-S, et al. The increase of voltage-gated potassium channel Kv3.4 mRNA expression in oral squamous cell carcinoma. J Oral Pathol Med 2003;32:606–11.

[132] Leanza L, Venturini E, Kadow S, Carpinteiro A, Gulbins E, Becker KA. Targeting a mitochondrial potassium channel to fight cancer. Cell Calcium 2015;58. doi:10.1016/j.ceca.2014.09.006.

[133] Lew T-S, Chang C-S, Fang K-P, Chen C-Y, Chen C-H, Lin S-C. The involvement of K(v)3.4 voltage-gated potassium channel in the growth of an oral squamous cell carcinoma cell line. J Oral Pathol Med 2004;33:543–9. doi:10.1111/j.1600-0714.2004.00236.x.

[134] Chung H-Y, Park YK. Rationale, Feasibility and Acceptability of Ketogenic Diet for Cancer Treatment. J Cancer Prev 2017;22:127–34. doi:10.15430/JCP.2017.22.3.127.

[135] Lv M, Zhu X, Wang H, Wang F, Guan W. Roles of caloric restriction, ketogenic diet and intermittent fasting during initiation, progression and metastasis of cancer in animal models: A systematic review and meta-analysis. PLoS One 2014;9. doi:10.1371/journal.pone.0115147.

[136] NEBELING LC, LERNER E. Implementing A Ketogenic Diet Based on Medium-chain Triglyceride Oil in Pediatric Patients with Cancer. J Am Diet Assoc 1995;95:693–7. doi:10.1016/S0002-8223(95)00189-1.

[137] Allen BG, Bhatia SK, Anderson CM, Eichenberger-Gilmore JM, Sibenaller ZA, Mapuskar KA, et al. Ketogenic diets as an adjuvant cancer therapy: History and potential mechanism. Redox Biol 2014;2:963–70. doi:10.1016/j.redox.2014.08.002.

[138] Pannone G, Bufo P, Santoro A, Franco R, Aquino G, Longo F, et al. WNT pathway in oral cancer: Epigenetic inactivation of WNT-inhibitors. Oncol Rep

2010;24:1035–41. doi:10.3892/or-00000952.

[139] Anastas JN, Moon RT. WNT signalling pathways as therapeutic targets in cancer. Nat Rev Cancer 2012;13:11–26. doi:10.1038/nrc3419.

[140] Barker N, Clevers H. Mining the Wnt pathway for cancer therapeutics. Nat Rev Drug Discov 2006;5:997–1014. doi:10.1038/nrd2154.

[141] Qin JJ, Nag S, Wang W, Zhou J, Zhang WD, Wang H, et al. NFAT as cancer target: Mission possible? Biochim Biophys Acta - Rev Cancer 2014;1846:297–311. doi:10.1016/j.bbcan.2014.07.009.

[142] Mancini M, Toker A. NFAT proteins: Emerging roles in cancer progression. Nat Rev Cancer 2009;9:810–20. doi:10.1038/nrc2735.

[143] Müller MR, Rao A. NFAT, immunity and cancer: A transcription factor comes of age. Nat Rev Immunol 2010;10:645–56. doi:10.1038/nri2818.

[144] Yeung N, Cline MS, Kuchinsky A, Smoot ME, Bader GD. Exploring biological networks with cytoscape software. Curr Protoc Bioinforma 2008. doi:10.1002/0471250953.bi0813s23.

[145] Manual CU. Cytoscape User Manual. Syst Biol (Stevenage) 2011;163:18–28. doi:10.1111/j.1476-5381.2010.01178.x.

[146] Watanabe K, Biesinger J, Salmans ML, Roberts BS, Arthur WT, Cleary M, et al. Integrative ChIP-seq/microarray analysis identifies a CTNNB1 target signature enriched in intestinal stem cells and colon cancer. PLoS One 2014;9. doi:10.1371/journal.pone.0092317.

[147] Geyer FC, Lacroix-Triki M, Savage K, Arnedos M, Lambros MB, MacKay A, et al. B-Catenin pathway activation in breast cancer is associated with triple-negative

phenotype but not with CTNNB1 mutation. Mod Pathol 2011;24:209–31. doi:10.1038/modpathol.2010.205.

[148] Morikawa T, Kuchiba A, Lochhead P, Nishihara R, Yamauchi M, Imamura Y, et al. Prospective analysis of body mass index, physical activity, and colorectal cancer risk associated with β-catenin (CTNNB1) status. Cancer Res 2013;73:1600–10. doi:10.1158/0008-5472.CAN-12-2276.

[149] Tornesello ML, Buonaguro L, Tatangelo F, Botti G, Izzo F, Buonaguro FM. Mutations in TP53, CTNNB1 and PIK3CA genes in hepatocellular carcinoma associated with hepatitis B and hepatitis C virus infections. Genomics 2013;102:74–83. doi:10.1016/j.ygeno.2013.04.001.

[150] Hirata H, Hinoda Y, Ueno K, Shahryari V, Tabatabai L, Dahiya R. MicroRNA-1826 targets VEGFC, beta-catenin (CTNNB1) and MEK1 (MAP2K1) in human bladder cancer. Carcinogenesis 2012;33:41–8. doi:10.1093/carcin/bgr239.

[151] Recurrent BTK and PLCG2 mutations confer ibrutinib resistance. Cancer Discov 2014;4:866. doi:10.1158/2159-8290.CD-RW2014-128.

[152] Stanislaus A, Bakhtiar A, Salleh D, Tiash S, Fatemian T, Hossain S, et al. Knockdown of PLC-gamma-2 and calmodulin 1 genes sensitizes human cervical adenocarcinoma cells to doxorubicin and paclitaxel. Cancer Cell Int 2012;12. doi:10.1186/1475-2867-12-30.

[153] Choi KY, Cho YJ, Kim JS, Ahn YH, Hong SH. SHC1 sensitizes cancer cells to the 8-Cl-cAMP treatment. Biochem Biophys Res Commun 2015;463:673–8. doi:10.1016/j.bbrc.2015.05.123.

[154] Zheng Y, Zhang C, Croucher DR, Soliman MA, St-Denis N, Pasculescu A, et al.

Temporal regulation of EGF signalling networks by the scaffold protein Shc1. Nature 2013;499:166–71. doi:10.1038/nature12308.

[155] SHC1 temporally regulates EGFR signaling. Cancer Discov 2013;3. doi:10.1158/2159-8290.CD-RW2013-154.

[156] Hoeller D, Hecker CM, Dikic I. Ubiquitin and ubiquitin-like proteins in cancer pathogenesis. Nat. Rev. Cancer, vol. 6, 2006, p. 776–88. doi:10.1038/nrc1994.

[157] Kirkin V, Dikic I. Ubiquitin networks in cancer. Curr Opin Genet Dev 2011;21:21–8. doi:10.1016/j.gde.2010.10.004.

[158] Ohta T, Fukuda M. Ubiquitin and breast cancer. Oncogene 2004;23:2079–88. doi:10.1038/sj.onc.1207371.

[159] Sun Y. Targeting E3 Ubiquitin Ligases for Cancer Therapy. Cancer Biol Ther 2003;2:621–7. doi:10.4161/cbt.2.6.677.

[160] Shi D, Grossman SR. Ubiquitin becomes ubiquitous in cancer: Emerging roles of ubiquitin ligases and deubiquitinases in tumorigenesis and as therapeutic targets. Cancer Biol Ther 2010;10:737–47. doi:10.4161/cbt.10.8.13417.

[161] Sebban S, Farago M, Gashai D, Ilan L, Pikarsky E, Ben-Porath I, et al. Vav1 Fine Tunes p53 Control of Apoptosis versus Proliferation in Breast Cancer. PLoS One 2013;8. doi:10.1371/journal.pone.0054321.

[162] Ilan L, Katzav S. Human Vav1 expression in hematopoietic and cancer cell lines is regulated by c-Myb and by CpG methylation. PLoS One 2012;7. doi:10.1371/journal.pone.0029939.

[163] Razidlo GL, Schroeder B, Chen J, Billadeau DD, McNiven MA. Vav1 as a central regulator of invadopodia assembly. Curr Biol 2014;24:86–93.

doi:10.1016/j.cub.2013.11.013.

[164] Razidlo GL, Magnine C, Sletten AC, Hurley RM, Almada LL, Fernandez-Zapico ME, et al. Targeting pancreatic cancer metastasis by inhibition of Vav1, a Driver of Tumor Cell Invasion. Cancer Res 2015;75:2907–15. doi:10.1158/0008-5472.CAN-14-3103.

[165] Nitsche M, Christiansen H, Hermann RM, Lücke EM, Peters K, Rave-Fränk M, et al. The combined effect of fludarabine monophosphate and radiation as well as gemcitabine and radiation on squamous carcinoma tumor cell lines in vitro. Int J Radiat Biol 2008;84:643–57. doi:10.1080/09553000802241754.

[166] Neves SS, Sarmento-Ribeiro AB, Simões SP, Pedroso de Lima MC. Transfection of oral cancer cells mediated by transferrin-associated lipoplexes: mechanisms of cell death induced by herpes simplex virus thymidine kinase/ganciclovir therapy. Biochim Biophys Acta 2006;1758:1703–12. doi:10.1016/j.bbamem.2006.08.021.

[167] Juneja M, Kobelt D, Walther W, Voss C, Smith J, Specker E, et al. Statin and rottlerin small-molecule inhibitors restrict colon cancer progression and metastasis via MACC1. PLoS Biol 2017;15. doi:10.1371/journal.pbio.2000784.

[168] Yin X, Zhang Y, Su J, Hou Y, Wang L, Ye X, et al. Rottlerin exerts its anti-tumor activity through inhibition of Skp2 in breast cancer cells. Oncotarget 2016;7:66512–24. doi:10.18632/oncotarget.11614.

[169] Soltoff SP. Rottlerin: an inappropriate and ineffective inhibitor of PKCδ. Trends Pharmacol Sci 2007;28:453–8. doi:10.1016/j.tips.2007.07.003.

[170] Bisson J, McAlpine JB, Friesen JB, Chen SN, Graham J, Pauli GF. Can Invalid Bioactives Undermine Natural Product-Based Drug Discovery? J Med Chem

2016;59:1671–90. doi:10.1021/acs.jmedchem.5b01009.

[171] Saloura V, Langerman A, Rudra S, Chin R, Cohen EEW. Multidisciplinary Care of the Patient with Head and Neck Cancer. Surg Oncol Clin N Am 2013;22:179–215. doi:10.1016/j.soc.2012.12.001.

[172] Marur S, Forastiere AA. Head and Neck Squamous Cell Carcinoma: Update on Epidemiology, Diagnosis, and Treatment. Mayo Clin Proc 2016;91:386–96. doi:10.1016/j.mayocp.2015.12.017.

[173] Mesía Nin R, Pastor Borgoñón M, Cruz Hernández JJ, Isla Casado D. SEOM clinical guidelines for the treatment of head and neck cancer. Clin Transl Oncol 2010;12:742–8. doi:10.1007/s12094-010-0589-2.