THE DEVELOPMENT OF A NEW PROGNOSTIC LOGISTIC REGRESSION MODEL: A CASE STUDY ON ORAL SQUAMOUS CELL CARCINOMA

MUHAMMAD AZEEM

UNIVERSITI SAINS MALAYSIA

2023

THE DEVELOPMENT OF A NEW PROGNOSTIC LOGISTIC REGRESSION MODEL: A CASE STUDY ON ORAL SQUAMOUS CELL CARCINOMA

by

MUHAMMAD AZEEM

Thesis submitted in fulfilment of the requirement
of degree of
Doctor of Philosophy

July 2023

ACKNOWLEDGEMENT

In the name of Allah, who is Lord of worlds and deserves all praise, He bestowed His favours abundantly on us.

Foremost, I would like to thank my main supervisor Associate Professor Ts. Dr. Wan Muhamad Amir for his contributions of time, ideas, and guidance, in making my research experience as innovative, productive, and stimulating. I appreciate his patience, proficiency, and expertise throughout the project. I am thankful for his supervision. I really appreciate the dedication and motivation of my co-supervisor Dr. Tang Lisen to complete our project. I would like to thank to USM to provide me USM fellowship.

I gratefully acknowledge my best friend Dr. Irfan Ullah Khan for his valuable advice, constant support, and crucial contribution to this thesis. I am really blessed to have him as a brother in my life. I would like to thank Dr. Mohammad Zahoor ul Haq and Dr. Farooq for their support and affectioning advice. My parents are the reason for all of my happiness. I can't begin to put into words how grateful I am to my parents for their never-ending love, support, guidance, and prayers.

I am unable to adequately convey my love for my dear siblings, who made many sacrifices by avoiding me for the majority of the time while I was studying here. Last but not least, I want to express my gratitude to my wonderful wife Saira, who has always supported me and my beautiful children Fatima ul Zahra, Muhammad Ismayial, Mustafa Azeem, and Muhammad Ahmad whose efforts and unwavering love helped me grow into the person I am today. This thesis is dedicated to my family.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

α	Alpha
Ac	Acetylation
ASR(W)	Age-standardized to world population
ANN	Artificial neural network
BPV-1	Bovine papillomavirus protein
CBP	CREB binding protein
CDK2	Cyclin-dependent kinase 2
CI	Confidence interval
DAB	3 - Diaminobenzidine
DNA	Deoxyribonucleic acid
EGFR	Epidermal growth factor receptor
ET-1	Endothelin-1
E6AP	E6-associated protein
FADD	FAS-associated protein with death domain
HPV	Human papillomavirus
HNSCC	Head and neck squamous cell carcinoma
HIV	Human immunodeficiency virus
HNC	Head and neck cancers
HR-HPV	High-risk human papillomavirus
Hospital USM	Hospital Universiti Sains Malaysia
HDACs	Histone deacetylases
I	Intensity
ISH	In situ hybridization
IARC	International Agency for Research on Cancer
IL-8	Interleukin-8

IHC	Immunohistochemistry
LCR	Long control region
MLFFNN	Multilayer feedforward neural network
MSE	Mean square error
MAD	Mean absolute deviance
MMPs	Matrix metalloproteinases
NGF	Nerve growth factor
OSCC	Oral squamous cell carcinomas
OED	Oral epithelial dysplasia
OR	Odds ratio
OPMDs	Oral potentially malignant disorders
ORFs	Open reading frames
OPSCC	Oral pharyngeal squamous cell carcinoma
PI	Pattern of invasion
pRb	Protein retinoblastoma
PAR_2	Protease-activated receptor type 2
PCR	Polymerase chain reaction
ROC	Receiver operation characteristics
SPSS	Statistical Package for the Social Sciences
TIF	Tumour invasion front
TNM	Tumour-node-metastasis
TGF-β	Transforming growth factor beta
TERT	Telomerase reverse transcriptase
USA	United States America
VEGF	Vascular endothelial growth factor
VIF	Variance inflation factor
WHO	World Health Organization

LIST OF APPENDICES

PEMBANGUNAN MODEL REGRESI LOGISTIK PROGNOSTIK BARU: KAJIAN KES TERHADAP KARSINOMA SEL SQUAMOUS ORAL

ABSTRAK

Oral Squamous Cell Carcinoma (OSCC) adalah kanser ke-enam paling kerap di seluruh dunia. Ia bertanggungjawab untuk 80-90% daripada semua neoplasma malignan mulut dan mempunyai kadar kematian sehingga 50%. Kanser mulut berpunca daripada etiologi pelbagai faktor, terutamanya merokok, tembakau, pengambilan alkohol, mengunyah sirih dan juga *human papillomavirus* berisiko tinggi. Seramai 57 pesakit dari klinik di Hospital Universiti Sains Malaysia telah dipilih sebagai sampel kajian. Dalam kajian retrospektif ini, teknik pengiraan bagi pemodelan statistik lanjutan digunakan untuk menilai deskripsi data pada tiga kajian bagi faktor yang berisiko: sosiodemografi, klinikopatologi dan ciri-ciri margin bagi pembedahan pesakit OSCC. Perisian R-Studio dan sintaks telah digunakan untuk mereka-bentuk, membangunkan kaedah biometri hibrid, mengaplikasikan dan juga menguji model yang telah dibina. Pendekatan lanjutan telah dilaksanakan dalam tiga bahagian, iaitu membangunkan sintaks R bagi kaedah hibrid biometry yang terdiri daripada metodologi bootstrap data, Rangkaian Neural Hadapan Suapan Berbilang Lapisan (MLFFNN) dan regresi logistik. Jantina lelaki, merokok, sirih kapur, dan tabiat pengambilan alkohol mempunyai hubungan yang signifikan dengan kematian (p <0.05). Antara ciri klinikopatologi yang meningkatkan saiz tumor, metastasis, OSCC yang sederhana dan kurang dibezakan, dan ekspresi Ki67 dikaitkan dengan pesakit yang telah meninggal dunia (p <0.05). Tambahan pula, ciri-ciri pembedahan radang margins perineural, radang tulang dan penglibatan pembedahan margin adalah signifikan terhadap kematian pesakit OSCC (p <0.05). Penemuan ini menyumbang kepada punca prognosis yang buruk. Kesimpulannya, terdapat potensi faktor yang

beririsiko berkaitan OSCC dalam populasi Malaysia. Kesimpulan kajian akan menggambarkan kehebatan teknik model hibrid yang digunakan dalam kajian.

THE DEVELOPMENT OF A NEW PROGNOSTIC LOGISTIC REGRESSION MODEL: A CASE STUDY ON ORAL SQUAMOUS CELL CARCINOMA

ABSTRACT

Oral squamous cell carcinoma (OSCC) is the sixth most frequent cancer worldwide. It is responsible for 80-90% of all mouth malignant neoplasms and has a mortality rate of up to 50%. Oral cancer has multifactorial etiology, mainly smoking, tobacco, alcohol consumption, betel quid chewing, and high-risk human papillomavirus (HPV). A total of 57 patients were recruited from the clinic at the Hospital Universiti Sains Malaysia (USM). In this retrospective study, advanced computational statistical modeling techniques were used to evaluate data descriptions on three risk factors studies: sociodemographic, clinicopathological, and surgical margins features of OSCC patients. The R-Studio software and syntax were used to design and develop the hybrid biometry method, implement, and the odd ratio. The advanced approach was executed in three parts, such as developing syntax for R for the biometry hybrid method which consists of data bootstrap methodology, multiple layer feedforward neural network (MLFFNN), and logistic regression. Male gender, smoking, betel quid, and alcohol habit variables were significantly related to death ($p < 0.05$). Among clinicopathological features increasing tumour size, metastasis, moderately and poorly differentiated OSCC, and Ki67 expression were significantly related to deceased patients ($p < 0.05$). Furthermore, features of surgical margins perineural invasion, bone invasion, and involvement of surgical margins were significantly related to the death of OSCC patients ($p < 0.05$). This finding might contribute to the underlying cause of poor prognosis. In conclusion, there exist

potential risk factors in relation to OSCC in the Malaysian population. The conclusion of the study might illustrate the superiority of the hybrid model technique used in the study.

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

In many countries, oral cancer is considered the leading cause of mortality due to oral diseases. Recent global estimates indicate that in 2018, there were 354,864 new cases and 177,384 fatalities (Bray, *et al.,* 2018)**.** Oral squamous cell carcinomas (OSCC) are the sixth most frequent malignant tumour (Ajila *et al.*, 2015), and is a fatal disease of the oral cavity with up to 50% of mortality rate (Mehrotra, *et al.*, 2006; Du, *et al.,* 2020). These cancers account for approximately 2 to 5% of all cancer cases worldwide, with Asia having the greatest prevalence (Siddiqui *et al*., 2006; Shield *et al*., 2017). Despite recent advances in treatment interventions in recent years, the survival rates has remained unchanged for decades (Mehrotra & Yadav, 2006). Oral malignancies have a multifactorial etiology, including smoking, tobacco utilization, alcohol intake, paan, betel quid, viral factors, and certain genetic and epigenetic alterations (Chaturvedi, 2012; Vargas-Ferreira *et al.*, 2012; Ali *et al.*, 2017; Jiang, *et al.*, 2019). According to the World Health Organization (WHO), the incidence of oral cancer is approximately age-standardized to the Malaysian community, and it is 3.0 per 100,000 individuals (Cheong *et al.,* 2017). The incidence of oral cancer is highest in Indian females where the ASR was 10.2/100,000 female populations. Deaths from oral cancer in Malaysia reached 1,587 in 2011, as reported by the Oral Cancer Research and Coordinating Center (OCRCC) in April 2011. At 7.72 deaths per 100,000 people over the age of 60, Malaysia is ranked fourteenth globally in terms of life expectancy (OCRCC 2017). In Kelantan, a previous study has shown that the 5-year survival of oral cancer patients was 18.0%, with a median survival time of 9

months. Poor survival was associated with being old, male, presenting in an advanced state, and not receiving treatment (Razak *et al*., 2010).

Tobacco use, heavy alcohol consumption, and the use of betel quid have all been linked to an increase in the incidence of oral cancer in many developing nations (Chaturvedi, 2012). Human papillomavirus (HPV) has been established as a significant causal factor in oral squamous cell carcinoma (OSCC) in recent studies. (Vargas-Ferreira *et al.*, 2012; Kerishnan *et al.*, 2016; Jiang & Dong, 2017)**.** Globally, Human papillomavirus is a major health crisis in a clinical setting. Tonsils and the bottom of the tongue are common locations for HPV-related cancers to manifest (Chaturvedi *et al.*, 2008).

The detection of HPV is also more common in subjects with more than one-lifetime sexual partner and with a history of oral sex. First-time researchers gave evidence of HPV in 1967 (Frithiof & Wersäll, 1967). For the detection of HPV, the first method had been introduced in 1982, which was antiserum to disrupt bovine papillomavirus protein (BPV-1) (Jenson *et al.*, 1982). Different methods have been used for the detection and genotyping of HPV. Among those, the most common methods being used are the detection of viral DNA with polymerase chain reaction (PCR), followed by In Situ Hybridization (ISH) and the detection of p16 by immunohistochemistry (Mirghani *et al.*, 2014). However, it is difficult to compare the result of HPV detection by different methods because of many reasons such as variation in sampling and site, selection of detection method of HPV, and the differences in method sensitivity (Krüger *et al*., 2014). In normal oral mucosa, the estimated prevalence of HPV ranges from 0.6% to 81% (D'Souza *et al.*, 2009; Ragin *et al.*, 2011). It has been documented in many studies oropharyngeal tumours often have HPV present.

Inaccurate assessment of oral cancer behavior may lead to improper management either as ineffective treatment or as unnecessary overtreatment. Therefore, identifying patients with low-risk or high-risk oral cancer can influence management decision-making and guide the selection of treatment approaches. Several prognostic markers have been suggested to improve the prognostication of oral cancer. At present, one of the most effective prognostic tools for tumour survival is the tumour-node-metastasis (TNM) staging system (Ribeiro *et al.*, 2000; Fan *et al.*, 2011; Almangush *et al.*, 2020). Moreover, patient's socio-demographic and clinical characteristics with their age, gender, and smoking habits are also considered for choosing the appropriate therapeutic strategy, to determine the risk of complications and the prognostic value of numerous distinct types of cancer (Rossi *et al.*, 2007; Kawakita *et al.*, 2012). These several identification factors are associated with a poor prognosis and thus have raised a critical issue in the treatment of OSCC. In previous studies, many different clinicopathological parameters such as age, smoking history, TNM staging (Kreppel *et al.*, 2013), tumour spread in cervical lymph nodes, tumour size and microvascular invasion (Grimm, 2012) have been studied as independent prognostic factors in patients with OSCC. It is necessary to mention that the previous studies on these prognostic factors have used traditional tools for data analysis, which have not produced any simple approach to utilize them as multiple prognostic factors should be applied to aid decision-making.

Machine learning, a subfield of Artificial intelligence has seen rapid growth in recent years, with much of this expansion attributable to the ever-increasing amount of medical data. The same can be said for artificial neural networks (ANNs), which are both fundamental to and a distinct area of machine learning. An ANN is a state-of-the-art hardware/software model whose operation is influenced by the brain. (Alhazmi

et al., 2021; Khanagar SB *et al.*, 2021). The complex relationship between input and output can be accurately modeled with comparatively simple computer programming code, further lending credence to the efficacy of ANN. Input, hidden, and output levels make up the ANN architecture (Al-Rawi *et al.*, 2022).

Furthermore, implement the validation by logistic regression. Logistic regression is a widely used method in machine learning for binary classification tasks, where the goal is to predict one of two possible outcomes (e.g., yes or no, true or false, positive or negative). The logistic regression model is trained by optimizing the parameters (weights) to minimize the difference between the predicted probabilities and the actual labels in the training data. This process is often done using maximum likelihood estimation or gradient descent-based optimization algorithms (Nusinovici, et al., 2020).

Nevertheless, these previous studies have evaluated only a limited number of risk factors that might affect the prognosis, when there is a possible dearth of information on the association between these factors with HPV and the prognosis of OSCC in Malaysia. Therefore, in this study, investigated the possible correlation between sociodemographic and clinicopathological factors and surgical specimen factors with treatment outcomes.

1.2 Problem Statement

Despite the advancement of treatment therapy, oral cancer prognosis is not improved by more than 50 % collective in the world. So, machine learning techniques may provide a viable alternative to traditional medical diagnosis or outcomes for oral cancer, as they are adept at dealing with noisy and incomplete data and allowing for

significant results to be achieved despite a small sample size. There were limited studies that focus on robust modeling and focusing on predictive modeling to give more accurate and precise results to make the prognosis of OSCC better. The purpose of this research was to investigate the demographics, clinicopathological characteristics, surgical reports, and treatment outcomes of patients with oral squamous cell carcinoma (OSCC) to improve prognosis.

1.3 Justification of the Study

Globally, oral cancer is the sixth most common malignancy that leads to a significant increase in the mortality rate. The prevalence of OSCC is common among Malaysian females (67.0%), and Indians (49.5%) majorly due to their habit of chewing betel quid (Kerishnan *et al.*, 2016). However, for Malays and Chinese groups, tobacco smoking and alcohol consumption are the significant risk factors (Zain *et al.*, 1999; Tan *et al*., 2000).

To date, there have been very scarce studies or no such studies have been published so far in the Malaysian population looking at Machine learning methods can handle noisy and incomplete data and produce significant results despite a small sample size, making them a potential alternative to traditional oral cancer diagnosis and prognosis of oral cancer. This study was undertaken to develop an advanced methodology of statistical analysis for the prognosis of oral cancer. Furthermore, this study aimed to investigate the sociodemographic, clinicopathological features, surgical report and treatment outcomes of patients who were diagnosed with OSCC.

There is an impending need to determine the treatment outcomes for patients treated in this hospital mainly to reflect on the current state of patient management and

treatment outcomes. There are many justified reasons why it is important to diagnose oral cancer in the early stage and refer for quick treatment as it will increase the 5-year survival rate.

Machine learning techniques are becoming popular in medical diagnoses and prognoses. Clinicopathologic traits determine prognosis. However, even the most skilled clinician cannot accurately prognosticate using these traits alone. Thus, forecast accuracy requires advanced statistical tools. This study uses feature selection and machine learning to predict oral cancer prognosis based on sociodemographic, clinicopathological, and surgical report traits.

Nevertheless, in Malaysia, there is still a lack of evidence indicating the necessity for these factors to be part of the management protocol of OSCC. As such we would like to provide more available data on this matter that reflects more on the local population in comparison to current international database information. This is important as it may act as baseline information to shed light on the future establishment of more effective management strategies, mainly targeting treatment protocols and patient education in local government hospitals.

1.4 Conceptional Framework of the Study

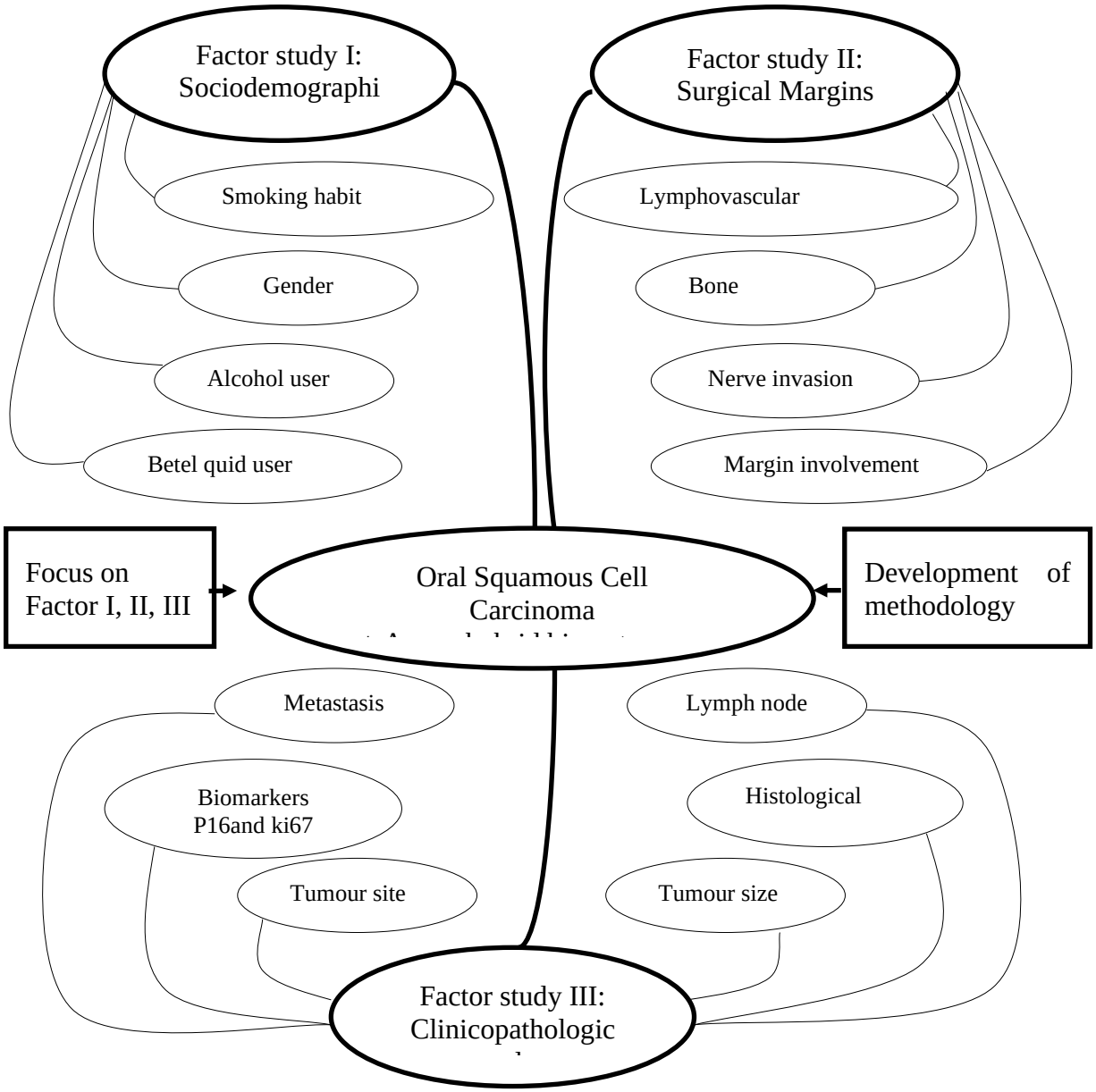

Figure 1.1 Conceptual Framework of the Study

1.5 Objectives of the Study

1.5.1 General Objective

To develop a new dimension for hybrid biometry prognostic cancer model in Oral Squamous Cell Carcinoma (OSCC).

1.5.2 Specific Objectives

a) To determine the socio-demographic, clinicopathological factors and surgical report features association with the outcome of treatment in a patient diagnosed with OSCC.

b) To design and develop a hybrid biometry method for OSCC through the R syntax using a combination of logistic regression and MLFFNN.

c) To implement hybrid biometric methods for modeling and prediction considering three different case studies of OSCC.

1.6 Research Questions

a) What are the significant sociodemographic, clinicopathological factors and surgical reports features that contribute to the treatment outcomes in patients who are diagnosed with OSCC?

b) How to develop a hybrid model for OSCC that yield efficient parameter estimates?

c) Is the hybrid model efficient for modeling and prediction purposes?

1.7 Research Hypothesis

There are significant associations of sociodemographic, clinicopathological factors and surgical reports features that contribute to the treatment outcomes in patients who are diagnosed with OSCC.

1.8 Organization of Thesis

This thesis consists of six distinct chapters. Chapter one discusses the study's background context, problem statement, objectives, research hypothesis, and conceptual framework. Chapter two provides a summary of the literature review on oral squamous cell carcinoma (OSCC), risk factors, incidence, detection methods, clinicopathological factors, and biomarkers multilayer feedforward neural network (MLFFNN). The third chapter discusses the study's design, inclusion and exclusion criteria, data capture method, development of R syntax, and analysis methodology. The results of the study on three distinct OSCC case studies are presented in Chapter 4. The results of case studies on the OSCC are discussed in chapter five, while chapter six contains the study's conclusion, strengths, and limitations. Moreover, it illuminate future recommendations.

CHAPTER 2

LITERATURE REVIEW

2.1 Oral Squamous Cell Carcinoma (OSCC)

According to National Cancer Institute (USA), oral cancers are defined as malignancies arising from the epithelial lining of tissues of the oral cavity or the oropharynx (NCI 2018). Squamous cell carcinomas account for the vast majority of these cancers (between 84% and 99%) (Ariyoshi et al., 2008; Kruaysawat et al., 2010), and they typically originate either from a preexisting "potentially malignant" area or, more frequently, out from ordinary epithelium. Oral squamous cell carcinoma (OSCC) is frequently used as a synonym for "oral malignancies" (Markopoulos, 2012). Lip, tongue, mouth (oral cavity), and oropharynx cancers are included in the global epidemiology of oral cavity and oropharynx cancers; however, cancers of the salivary glands and other pharyngeal locations are not (Warnakulasuriya, *et al.*, 2009; Du *et al.*, 2020).

2.2 Sites of OSCC

Over ninety percent of oral cavity tumours are OSCC, it has been said (Johnson et al., 2011; Markopoulos et al., 2012). Tumors originating in the lips, the front two-thirds of the tongue, the buccal mucosa, the hard palate, the floor of the mouth, the upper and lower alveolar ridges, the retromolar trigone, and the sublingual region are all included under the umbrella term "oral cancer" (Feller & Lemmer, 2012; Sundermann et al., 2017; Du et al., 2020). Nearly ninety-five percent of all malignant neoplasms are squamous cell carcinomas (Ayaz *et al.*, 2011; Jiang *et al.*, 2019).

2.3 Histology of OSCC

Multiple traits aid in the progression of OSCC, which results from a complex cascade of genetic alterations and epigenetic abnormalities in the signaling pathways that initiate cancer (Rivera & Venegas, 2014). Invasive epithelium neoplasms with varying degrees of squamous differentiation and keratinization are known as squamous cell carcinomas. It develops in either a keratinized (skin) or non-keratinized (oral mucosa, uterine exocervical mucosa, and oesophageal mucosa) stratified squamous epithelium (Suciu et al., 2014). Traditional OSCC, verrucous OSCC, spindle OSCC, basaloid OSCC, adenosquamous OSCC, papillary OSCC, and acantholytic OSCC are just some of the histopathological subtypes (Minhas et al., 2016). Smoking, alcohol intake, human papillomavirus, and other risk factors all contribute to the development of oral squamous cell carcinoma (Wilkey et al., 2009).

Oral squamous cell carcinoma can arise from several sites in the oral cavity but the most common sites are the lateral border of the tongue and floor of the mouth, with a percentage of 20–40% and 15–20%, respectively (Jerjes *et al.*, 2010; Feller and Lemmer, 2012). Histologically, the lesion ranges from various stages i.e. preneoplastic changes to the formation of cancer. However, it should not be considered that all reactional or precancerous lesions will result in the formation of a malignant tumour (Rivera & Venegas, 2014).

2.3.1 Potentially Malignant Changes in Oral Mucosa

The development of OSCC is majorly preceded by a series of specific histopathological changes. The term oral epithelial dysplasia (OED) is the diagnostic marker of premalignancy which is confined to the surface epithelial layer (Van der Waal, 2009).

Histopathological grading of OED can be used as a clinical tool to estimate the risk of malignant transformation and usually guides clinical management and treatment of patients (Dost *et al.*, 2014). These changes often manifest clinically as an oral mucosal lesion. During carcinogenesis, histological characteristics of epithelium can be classified as either reactive epithelial changes or neoplastic epithelial changes. Reactive epithelial changes include hyperplasia, hyperkeratosis, and acanthosis. Neoplastic epithelial changes include mild, moderate, and severe dysplasia (Wang *et al.*, 2009b). Oral cancer initially emerges as an epithelial dysplasia and was described as altered proliferation on the surface epithelium followed by degeneration of the subepithelial basement membrane, lending to local invasion and metastasis. Local invasion is characterized by the presence of cords and islets of epithelial cells within the underlying connective tissue (Fuentes *et al.*, 2012).

2.3.2 Histological Classification of Tumours

At present, to histologically classify the tumour lesions, two systems are used i.e. the international histological classification of tumours and the pattern of the tumour invasion front (TIF) (Rivera *et al.*, 2011). Initially, the classification is based on the degree of differentiation of a tumour which is categorized as well, moderate, and undifferentiated types (Dissanayaka *et al*., 2012). On the other hand, TIF is the most characteristic part of a malignant tumour and represents an area of the tumour lesion having the greatest depth of invasion and progression into the surrounding normal tissues (Wang *et al.*, 2009a). Moreover, the TIF cells show different molecular characteristics compared to the cells at the superficial surface of a tumour (Costa *et al*., 2015). Tumour invasion front is classified by four characteristic features which include the degree of keratinization, nuclear polymorphism, lymphocytic infiltrations,

and pattern of invasion (PI) (Sandu *et al*., 2014). Among these, PI is a useful prognostic factor in OSCC (Dissanayaka *et al.*, 2012). Different morphological standards are known to study the severity of invasion which ranges from islet-infiltrating cells with wide fronts of invasion, thin infiltrating cords, and individual infiltrating cells (Dissanayaka *et al.*, 2012).

2.4 Incidence of OSCC

It is widely acknowledged that mouth cancer is on the rise as a major health concern in countries all over the globe (Bray *et al.*, 2018). In 2020, lip and oral cavity cancer incidence counted 264,000 new cases in males (70% of total cases) and 113,000 in females, with a male/female ratio equal to 2.3:1; a prevalence in 5 years of over 656,000 and 303,000 and mortality for over 125,000 and 53,000, respectively. Among them, the cases attributable to alcohol were 67,000 for males and 8200 for women (Di Spirito, et al., 2022).

In South and Central Asia, oral squamous cell carcinoma ranks third most prevalent among all cancers (Warnakulasuriya, 2009; Jemal et al., 2014). Lip and oral cavity cancer is the 12th most prevalent cancer overall, with an incidence rate of 3.8% in South Asia, 1.8% in East Asia, and 2.1% in West Asia (GLOBOCAN, 2018). Oral cancer (including cancers of the mouth, oral cavity, and pharynx) is caused by rapidly growing tumours; the disease is more common in men than women (2:1) and occurs at a higher rate in South Asia (48.7%). (Shield et al., 2017). There were 14.1 million new cases of cancer and 8.2 million deaths from cancer in 2012, according to recent research. According to the World Health Organization (WHO), Southeast Asia has the highest incidence of oral cavity and lip cancer at 6.4 per 100,000. This is followed by Europe and the East Mediterranean (4.6 per 100000), the Americas (4.1 per 100000),

Africa (2.7 per 100000), and the Western Pacific area (2.0 per 100000). (Ferlay et al., 2013). Southern Asia, Eastern Europe, Latin America, and the Caribbean have the highest rates of oral squamous cell carcinoma worldwide (Warnakulasuriya, 2009).

Many geographic variations have been reported in the incidence of the lip and oral cavity cancer with the highest rate being reported from Malaysia, South Central Asia, and East and Central Europe, while the lowest rates are from West Africa and East Asia. According to OCRCC Malaysia, oral cancer is among the top 10 cancers in the Indian ethnicity. In recent years, the incidence rate has increased in men and women from Asia, North America, and Australia, and in men from South and West Europe because of the increasing tobacco use along with the higher prevalence of HPV infection in some countries (Yako-Suketomo & Matsuda, 2010; Torre *et al.*, 2015).

2.5 Risk factors for OSCC

There is a substantial difference in the global incidence of OSCC due to the indicative difference in exposure to environmental and behavioural risk factors. Oral cancer risk factors include tobacco use (chewing and smoking), alcohol intake, betel quid chewing, and persistent viral infections like HPV (Conway et al., 2018; Jiang et al., 2019). It is widely accepted that tobacco and alcohol play a significant role in oral cancer, but diet and oral hygiene are also being considered as possible separate risk factors (Zeng et al., 2013).

2.5.1 Major Risk factors for OSCC

One of the most significant causes of oral cavity cancer is still smoking. IARC (International Institute for Research on Cancer) ranks tobacco use as a group one carcinogen (the highest possible rating) for oral cavity cancer. Especially in developing

and middle-income nations, smoking has become increasingly common (WHO, 2022). In the absence of other risk factors, heavy smokers have a much greater chance of developing oral cancer than low smokers or nonsmokers. When a person smokes more than 20 cigarettes per day for more than 20 years, their chance of developing head and neck squamous cell carcinoma (HNSCC) increases dramatically (Applebaum, et al., 2007; Hashibe, et al., 2007; Kasza, et al., 2017).

After confirmation of the risk of carcinogenesis to tobacco smoking and other additional confounding factors, some studies in America, Europe, and Asia reported heavy alcohol consumption association with OSCC (WHO, 2015). A meta-analysis showed that chances of HNSCC increase with increasing doses of alcohol consumption. Those who use a high quantity of alcohol are at more risk to develop HNSCC as compared to those who use a low quantity of alcohol (Turati *et al.*, 2012; Kumar, *et al.*, 2016).

Heavy tobacco users have 20 times more risk; heavy alcohol users have a 5-fold higher risk and 50 times more risk for those who use both tobacco and alcohol. There is a synergistic effect between alcohol users and tobacco consumption (Mello, *et al.*, 2019).

Betel quid chewing is a frequent habit in various regions of Asia as well as Asian migrants worldwide (Petti *et al*., 2012; Mello, *et al.*, 2019).). The habit produces addictive psycho-stimulation with complaisant effects and is deeply rooted in many cultures (Petti *et al.*, 2012). The carcinogenic effect of betel quid is already established, and the carcinogenic effect is attributed to tobacco. The odds ratio (OR) in HNSCC for betel quid chewing with tobacco is 7 to 8, and 3 to 6 for betel quid without tobacco

(Guha *et al.*, 2014). Betel-quid chewing is also an important risk factor in Southeast Asia (Campisi *et al.*, 2007).

2.5.2 Minor Risk Factors for OSCC

Carcinogenesis may be influenced by not taking care of oral hygiene. Oral cancer has been linked to a high bacterial burden, which is common in conditions like chronic periodontitis and inadequate dental care, according to a few studies (Fitzpatrick and Katz, 2010; Gondivkar et al., 2013; Ahrens et al., 2014). Dietary practices may also play a role in oral cancer, though they are not as strong a risk factor as tobacco use or excessive sun exposure. A diet rich in fruits and veggies and low in meat has been proposed as a preventative measure against oral cancer by one author (Bravi et al., 2013). The vitamin D deficiency in oral cancer patients has been documented by some researchers (Verma, et al., 2020). However, the function of vitamin D in oral cancer requires more study.

The immune system plays an important role in carcinogenesis. Human Immunodeficiency Virus (HIV) infection and organ transplant patients have a high incidence of developing oral cancer as compared to the general population (Van Leeuwen *et al*., 2009; Collett *et al*., 2010). Environmental pollution might be playing a carcinogenic role as well as researchers had shown heavy carcinogenic element in the soil of Taiwan (Chiang *et al*., 2010). However, supporting data is inconsistent.

2.6 Pathophysiology of OSCC

The tumorigenic genomic alterations are of two major types i.e. tumour suppressor genes (p16 and p53), which when inactivated, stimulate tumour development through genetic mutation, loss of heterozygosity, or deletion, or by

epigenetic modifications such as remodeling of chromatin; and oncogenes (myc, erbB-2, Epidermal Growth Factor Receptor (EGFR), cyclin D1, which stimulate the growth of tumour upon activation through overexpression due to amplification, increased transcription, or changes in genetic structure (Huang *et al.,* 2006; Diez-Perez *et al.,* 2011). Both kinds play a significant role in the pathogenesis of OSCC. Alterations in p53 and p16 are related to the carcinogenesis process as p53 is responsible for regulating cell proliferation, DNA repair and apoptosis, whereas p16 regulates the cell cycle. Mutation in p53 is highly related to smoking and usage of tobacco in HNSCC (Hashibe *et al*., 2009).

2.6.1 Chemical Mediators

Endothelin-1 (ET-1), nerve growth factor (NGF) and proteases are involved in oral cancer. ET-1 is a vasoactive peptide that generates nociception by binding to endothelin-B receptors and signifies on dorsal root ganglion satellite cells and non-myelinating Schwann cells (Pickering *et al*., 2008; Schmidt, 2015). Protease-activated receptor type 2 (PAR_2) is also participating in oral cancer and this receptor is stimulated by serine proteases, tryptase, and trypsin (Russo *et al.,* 2009). Due to continuing action of serine proteases, sensory neurons of many microenvironments cancer exposed NGF (Jemal *et al.,* 2011). Nerve growth factor can promote the cell proliferation and invasion of oral cancer (Kolokythas *et al*., 2010).

2.6.2 Neovascularization

Angiogenesis is defined as “the growth of new blood vessels (neovascularization) from pre-existing ones, is a multi-step process, which appears to be regulated by both stimulatory and inhibitory factors and is vital for the continued

growth and survival of solid neoplasms" (Wadhwan *et al*., 2015). Angiogenesis is a key step in abnormal cell growth and metastasis of a tumour. Several different angiogenic factors have been involved i.e. vascular endothelial growth factor (VEGF), matrix metalloproteinases (MMPs), transforming growth factor beta (TGF-β), interleukin-8 (IL-8) (Grothey and Galanis, 2009; Deryugina and Quigley, 2015). Of these, expression of VEGF family, VEGF-A and VEGF-C are the important types that have been reported in OSCC and are majorly associated with metastasis (Friedrich *et al*., 2010; Okada *et al*., 2010; Kapoor and Deshmukh, 2012; Zhao *et al*., 2013).

Oral squamous cell carcinoma is a diversified process in which multiple genetic alterations occur that leads to the downregulation of tumour suppressor genes. These multiphasic events can result in alterations of angiogenic growth factors (Calixto *et al*., 2014).

2.7 Human Papillomavirus (HPV)

2.7.1 HPV Genomic Structure

Structurally, HPV has a closed circular double-stranded deoxyribonucleic acid (DNA) genome and this virus is characterized by a non-enveloped icosahedral capsid with a virion size of approximately 55 nm in diameter (Nemes *et al.*, 2006; Zheng and Baker, 2006). The DNA of HPV is around 8 kb in size which are divided into three major regions i.e. early (occupies over 50% of the virus genome), late (40% of the virus genome), and a long control region (LCR; 10% of the HPV genome). The molecular arrangement showed that all HPV have the same genomic organization with 8-9 open reading frames (ORFs) on corresponding strands. The gene expression of HPV is not fully understood however, it is believed that genomic expression leads to the expression of six viral non-structural regulatory proteins (E1, E2, E4, E5, E6 and

E7) from the first coding section late genes codes the major and minor capsid proteins L1 and L2, respectively, while the second coding section undergoes for terminal differentiation. The third section holds the LCR which is confined to ORF L1 and E6. All HPV genomes have different length of LCR (Fertey *et al*., 2011).

2.7.2 HPV Life Cycle

The differentiation of the epithelium that HPV infects is intrinsically linked to the reproduction cycle of HPV. An initial, low-multiplicity virus establishes itself in the epithelium's basal layer. The circular viral genome is transported into the cell, where it is uncoated and eventually delivered to the nucleus. After HPV has exited the cell, DNA replication will begin; this is the "active phase," and it is during this time that the virus replicates by making more than 1000 copies of its genome per cell. After that, the expression of late genes starts and finally viral particles are produced and released (Tommasino, 2014). (Tommasino, 2014). Even though the oncoproteins, E6 and E7 are also expressed shortly after infection, it is difficult to assess the precise temporal lineage of initial events in the virus life cycle. The complexity of HPV infection and tracking virus progression is a major factor. On the other hand, the infection may linger for a prolonged period of time, even for several years, within the infected cell, its daughter cells, and the basal layer of the epithelium (McKinney et al., 2015).

2.7.3 HPV Subtypes and Diseases

Human papillomavirus is one of the major causative agents in the disease of the skin and mucosal epithelia. The diseases range from common benign warts of the hands to potentially invasive cervical cancer. Human papillomavirus has over 170

subtypes which are divided broadly into two groups as low-risk (e.g. types 6 and 11) and high-risk (e.g. types 16 and 18) (Ghittoni *et al.*, 2015). Low-risk HPV can cause genital warts, while high-risk HPV is associated with malignant lesions such as cervical neoplasm, penile, oropharyngeal, and vulvar carcinomas (Viens, 2016).

2.7.4 HPV as a Predisposing Risk Factor in OSCC

Recent research has pointed to HPV infection as a significant risk factor for OSCC, joining tobacco use and alcohol consumption as known aetiologies. Oral HPV infection is commonly spread through sexual practices like open-mouth kissing (D'souza et al., 2007; Gillison et al., 2008; Heck et al., 2010). The researcher's Gan et al. discovered that the prevalence of HPV detection was higher in tobacco and tobacco product users compared to non-users (Gan et al., 2014). In patients who have had more than one sexual partner in their lives, and who are willing to discuss their sexual history openly, the detection rate is higher. Based on IARC's previous assessment of HPV, the latest estimate places the percentage of cancers originating in the mouth and throat at 25.6%, with HPV 16 showing the highest persistence (Bruni *et al.*, 2017). What's more, HPV 16 and 18 are firmly linked to virtually all HPV-positive cancers (Kreimer, 2014).

Within the HPV family, there is more prevalence and a significant association with OSCC among high-risk HPV types (HPV 16 and 18) compared to other low-risk types. The probability of detecting high-risk HPV in OSCC was estimated to be 2.8 times more than that of low-risk HPV (Miller and Johnstone, 2001). In a study by Zhang et al., it has been found that 74% of OSCC were positive for HPV 16 and 18 DNA compared to the normal oral mucosa, which is reported to be fifty-five per cent (Zhang *et al.*, 2004). In a case-control study conducted in the Mexican population, strong evidence between HPV and OSCC has been observed. Moreover, it has been

found that the prevalence of HPV was more in OSCC cases (43.5%) compared to the control (17.3%) with the most frequent types were HPV 16 and 18 among OSCC cases (Anaya-Saavedra *et al.*, 2008). A meta-analysis by Miller et al. on HPV has shown that the possibility of detecting HPV in the oral mucosa of OSCC patients is 46.5% compared to HPV in normal oral mucosa which is 10% (Miller and Johnstone, 2001). Few studies on the Indian population have also shown a similar prevalence rate of HPV in OSCC patients as 70.59% in Karnataka, 45% in an unknown place in India, and 43.54% in Northeast India (Kulkarni *et al.*, 2011; Jalouli *et al.*, 2012; Mondal *et al.*, 2013).

Besides these, there are few published studies on Australian (6%), Indian (0% to 1.6%), Malaysian (3.3%) and German (6%) populations which have also shown low to zero prevalence of HPV in OSCC (Goot-Heah *et al.*, 2012; Krüger *et al.*, 2014; Antonsson *et al.*, 2015; Chen *et al.*, 2016a; Gheit *et al.*, 2017).

Based on these studies, it is believed that there is a strong association between HPV and HPV-related OSCC. The high-risk HPV types possibly have a more significant association with OSCC, compared to low-risk types. Moreover, the incidence of HPV-positive OSCC has increased significantly in recent decades.

There is a tragic increase in HPV associated HNSCC in the world with some variations according to gender, ethnicity, and geographical location. HPV is an etiological factor, has been established recently by the International Agency for Research on Cancer (IARC) for oropharyngeal squamous cell carcinoma (Dalianis, 2014). The prevalence of HPV-related head and neck cancers (HNC), especially oropharyngeal cancer is more in North America (70%) and Europe (50%) when compared to rest of the world (Stein *et al.*, 2015). A rising trend has been seen in

Australia, where HPV-positive oral pharyngeal squamous cell carcinoma (OPSCC) gradually raised from twenty percent to sixty-three percent of cases over the recent two decades (Hong *et al.,* 2016). A recent study on Bangladeshi population has shown 21% high-risk HPV infection in HNSCC (Chowdhury *et al.,* 2017). Furthermore, a study in India has shown high-risk HPV detection (13.7%) in HNC (Gheit *et al.*, 2017). A meta-analysis from 2002 to 2012 performed on the European population has shown that the prevalence of high-risk HPV is forty percent in HNC (Abogunrin *et al.*, 2014). A meta-analysis by Ndiaye et al. included 148 studies from 44 different countries and has shown 31.5% HPV detection in HNSCC (Ndiaye *et al.*, 2014). There is an increased incidence of HPV-related HNC in Taiwan from 1995 to 2009 which has been found in most of the younger patients as compared to non-HPV-associated HNC. The overall incidence of HNSCC is also rising in Taiwan and this is an alarming issue (Hwang *et al.*, 2015).

It has been suggested that the increased incidence of HPV is related to early sexual exposure, high numbers of sexual partners and oral sex habits (D'Souza *et al.*, 2009). Nowadays, HPV status is very important prior to treatment planning because HPV-associated HNSCC patients have a better prognoses regardless of the tumour stage, age, and gender (Ang *et al.,* 2010; Bonilla-Velez *et al.*, 2013).

2.7.5 Pathogenesis of HPV-induced OSCC

The HPV DNA consists of 7200–8000 base pairs which are organized into three regions i.e. early region, the late region, and the genomic regulatory region. The early region comprises of E1, E2, E4–E7 and characterizes 50%, the late region comprising of L1 and L2 which represents 40% and the genomic regulatory region represents 10% of the genome. E6 and E7 oncoproteins have a very significant role in

the malignant transformation of infected cells by downregulation of tumour suppressors gene p53 and protein retinoblastoma (pRb) (Leemans *et al.*, 2011; Tommasino, 2014). Other studies have shown that E6 and E7 play a vital role in the proliferation of epithelium in a benign and malignant tumour related to HPV (Boscolo-Rizzo *et al.*, 2013).

2.8 Favourable Response to Therapy of HPV Positive Head and Neck Squamous Cell Carcinoma (HNSCC)

There is a significant finding that indicates a better prognosis in HPV-positive oropharyngeal cancer. This statement was supported by a study that showed that patients with HPV-positive oropharyngeal cancers with p16 overexpression had a considerably better overall survival ($p = 0.01$) and disease-specific survival ($p = 0.046$) rate as compared to non-HPV or non-p16 expression patients (Cai *et al*., 2014). Another study showed that improved response rates after chemotherapy (82% vs 55%) and after induction chemoradiation treatment (84% vs 57%) in HPV positive tumours (Fakhry *et al.*, 2008).

The reasons for improved prognosis in HPV-related cancers remains hypothetical. Some studies supported the theory of improved radiation response towards viral antigen (Lassen *et al*., 2009; Rieckmann *et al*., 2013). There is an evidence which shows that quick regression following response to radiation in an HPV-related tumour as compared to HPV a negative tumour (Chen *et al*., 2013).

2.9 HPV and Radio-Sensitivity

In most cases of HNSCC, radiation is the primary treatment. HPV-related tumours which have decreased rates of p53 mutation may have higher radiosensitivity and have shown better initial regression of a tumour (Chen *et al*., 2013). One experimental study had shown HPV positive cell line have higher radiosensitivity compared to negative cell lines. However, compromised DNA repair capability is likely to acquit an improved outcome (Rieckmann *et al.*, 2013).

As proposed by clinical data, a more favourable response to radiation, chemotherapy or surgery is possible in HPV-positive tumours (Reimers *et al.*, 2007). A meta-analysis had shown that a favourable prognosis is seen in OPSCC with p16 expression carcinomas (Sedghizadeh *et al.*, 2016).

2.10 Clinico Pathological Prognosticators in OSCC

Despite major advances in diagnostics and therapeutics, the overall survival rate remains between 45% and 50% in many nations, with no evidence of improvement over the past few years (Omar, 2013). Presently, the prognosis and treatment strategy for people with oral cancer are based primarily on clinical and histopathological factors.

There are certain pathological prognosticators implicated in OSCC which includes tumour site, tumour size, histological grading, perineural invasion, lymphovascular invasion, and bone and cervical lymph nodes involvement (Woolgar, *et al.*, 2009; Dolens, *et al.*, 2021) . The associations between clinical characteristics of OSCC patients and a poor prognosis were assessed. The resultant analysis revealed that male gender ($p = 0.123$), a history of smoking ($p = 0.225$), tumour stages of I and