Community Dental Health (2023) **40,** 252–260
Received 27 January 2023; Accepted 7 June 2023

© BASCD 2023
doi:10.1922/CDH_00015Ghanati09

Methodological Issues with Head and Neck Cancer Prognostic Risk Prediction Models

Hamed Ghanati,[1] Sreenath Madathil,[1] Mohammad Al-Tamimi,[1] Ziad Al Asmar,[1] Martin Morris[2] and Belinda Nicolau[1]

[1]Faculty of Dental Medicine and Oral Health Sciences, McGill University, Canada; [2]Schulich Library of Physical Sciences, Life Sciences and Engineering, McGill University, Canada

Objective: Prognostic risk prediction models estimate the probability of developing head and neck cancer (HNC), providing valuable information for managing the disease. While different prognostic HNC risk prediction models have been developed worldwide, a comprehensive evaluation of their methods is lacking. We conducted a scoping review with a critical assessment aiming to identify the methodological strengths and limitations of HNC risk prediction models. ***Method***: We searched Medline, Embase, Scopus, Web of Science, and CAB Abstracts databases and included full-text-available peer-reviewed published papers on developing or validating a prognostic HNC risk prediction model. Study quality was appraised using the PROBAST tool. ***Results***: Nine papers were included. Although all had a high risk of bias, mainly in the analysis domain, only two studies had high concerns about clinical applicability. ***Conclusion***: Currently published studies provide insufficient information on methods, making it difficult to judge the models' quality and applicability. Future investigations should follow the guidelines in reporting the prediction modelling studies.

Keywords*: Statistical, Prognosis, Models, Risk Assessment, Head and Neck Neoplasms, Review*

Introduction

Every year, more than 700,000 cancers of the lips and oral cavity, oropharynx, hypopharynx, and larynx, also known as Head and Neck Cancers (HNCs), are diagnosed around the world (Sung *et al.*, 2021). Due to their anatomic location, HNCs have one of the highest morbidity rates, with a 5-year survival rate of around 50% (Tiwana *et al.*, 2014). Tobacco smoking, alcohol consumption, and human papillomavirus (HPV) are the main risk factors for HNC (Dhull *et al.*, 2018; Toporcov *et al.*, 2015). Despite this knowledge, their incidence has remained relatively stable (Carvalho *et al.*, 2005; Johnson-Obaseki *et al.*, 2012; Marur and Forastiere, 2008). Importantly, the incidence of a subset of HNC related to HPV is increasing in several countries (Argirion *et al.*, 2019; Joseph and D'Souza, 2012), specifically in developed countries (Curado and Hashibe, 2009) (Habbous *et al.*, 2017; Johnson-Obaseki *et al.*, 2012). There is, therefore, a need to devise prevention strategies to reduce HNC incidence (Hashim *et al.*, 2019).

Risk prediction models have become increasingly popular in medical decision-making (Chen, 2020; Shipe *et al.*, 2019; Steyerberg, 2019). These models estimate the probability of having a disease (diagnostic prediction model) or future occurrence of a disease (prognostic prediction model) based on an individual's sociodemographic and behavioral characteristics (Hendriksen *et al.*, 2013; Steyerberg, 2019). Thus, they may assist healthcare professionals (Domchek *et al.*, 2003) with personalized prevention intervention strategies (Silveira *et al.*, 2018). Prediction models are also helpful in identifying high-risk individuals for screening programs (Tammemaegi, 2015) or clinical trials for new prevention measures (Chatterjee *et al.*, 2016). Notably, they have been successfully applied in lung cancer screening programs (Guo *et al.*, 2022; Tammemaegi, 2015; Tammemägi *et al.*, 2022). We can, therefore, expect that these models will provide similar assistance in identifying high-risk individuals for HNC screening programs (Cheung *et al.*, 2021) and clinical trials (e.g., trials to prevent oral HPV infection) (Diana and Corica, 2021).

Multiple prognostic risk prediction models have been developed for HNC (Cheung *et al.*, 2021; Gupta *et al.*, 2017; Hung *et al.*, 2020; Koyanagi *et al.*, 2017; Krishna Rao *et al.*, 2016; Lee *et al.*, 2020; McCarthy *et al.*, 2020; Rosma, 2010; Tota *et al.*, 2019). However, there is limited evidence regarding the quality of their methodology, raising concerns about their applicability in clinical settings, public health, and clinical trials. A recent rapid review assessed the quality and clinical applicability of various HNC risk prediction models (Smith *et al.*, 2022). However, this work did not provide a critical and comprehensive analytical assessment of these models.

This scoping review systematically maps the literature on HNC prognostic risk prediction modelling, employing a methodological lens to thoroughly evaluate the performance, risk of bias, and practical applicability of the existing models. By investigating the methodological strengths and limitations of the current models, this review will identify the most well-developed and reliable models, providing valuable insights for the future development of HNC prediction models.

Correspondence to: Prof. Belinda Nicolau. Email: belinda.nicolau@mcgill.ca

Method

Following a thorough literature review, our team formulated the research question: “What are the methodological concerns associated with existing prognostic HNC risk prediction models?” This question investigated the study designs, data sources, model types and strategies currently used in the literature to develop HNC prognostic risk prediction models. We used the Population, Concepts, and Context (PCC) framework to define the research question (Westphaln *et al.*, 2021). The population of interest was any type of prognostic model developed to predict the individual risk of developing HNC. The concept was the model development, validation strategy, and performance metrics. The context comprised studies developing or validating at least one HNC prognostic risk prediction model. We followed an updated scoping review methodology of Arksey and O’Malley (Arksey and O’Malley, 2005; Westphaln *et al.*, 2021) proposed by Levac et al. (2010).

Information source and literature search

A medical librarian (MM) created a systematic scoping search strategy (Morris *et al.*, 2016) for Medline (Ovid), comprising a combination of Medical Subject Headings, title/abstract keywords, truncations, adjacency operators, and Boolean operators and included the concepts of head and neck cancers, epidemiology, and computer modelling (Supplementary Table I available at https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP3/AV7K47). The strategy was subsequently translated for Embase (via Ovid), CAB Abstracts, Scopus, and Web of Science. All databases were searched from inception to 18 June 2021, and the combined library was deduplicated in Endnote 20 (Gotschall, 2021). We scanned the reference lists of included articles to find any missed publications. Two blinded reviewers (HG & ZA) shortlisted the papers on Rayyan (Ouzzani *et al.*, 2016). We included peer-reviewed full-text-available papers and those papers developing or validating at least one HNC prognostic risk prediction model. Review articles and papers that discussed genetic predictors (e.g., DNA methylated genes as predictors) were excluded as we focused models on models applicable to clinical settings. Reviewers’ conflicts were resolved by discussing with two experts (SM & BN). The inter-reviewer agreement was assessed using Cohen’s Kappa coefficient (Cohen, 1960). Two investigators (HG and MA) extracted the data based on TRIPOD criteria(Collins *et al.*, 2015). The quality of the studies was further evaluated using the PROBAST (Moons *et al.*, 2019; Wolff *et al.*, 2019), and the results were reported based on the PRISMA-ScR (Peters *et al.*, 2020).

Results

Our search strategy found 1554 articles, of which 192 were duplicates (Figure 1). Nine papers met the inclusion criteria (Cheung *et al.*, 2021; Gupta *et al.*, 2017; Hung *et al.*, 2020; Koyanagi *et al.*, 2017; Krishna Rao *et al.*, 2016; Lee *et al.*, 2020; McCarthy *et al.*, 2020; Rosma, 2010; Tota *et al.*, 2019). The Kappa coefficient was 75.34%, indicating good inter-reviewer agreement. Table 1 presents the characteristics of the included studies. Four papers were from the last three years (Cheung *et al.*, 2021; Hung *et al.*, 2020; Lee *et al.*, 2020; McCarthy *et al.*, 2020), while five were published between 2010 and 2019 (Gupta *et al.*, 2017; Koyanagi *et al.*, 2017; Krishna Rao *et al.*, 2016; Tota *et al.*, 2019). Data sources included population-based cohorts (Hung *et al.*, 2020), case-controls (Gupta *et al.*, 2017; Koyanagi *et al.*, 2017; Krishna Rao *et al.*, 2016; Lee *et al.*, 2020; McCarthy *et al.*, 2020; Rosma, 2010; Tota *et al.*, 2019), and randomized controlled screening trial (Cheung *et al.*, 2021). Most studies (60.0%) used data from Asian countries such as India (Cheung *et al.*, 2021; Gupta *et al.*, 2017; Krishna Rao *et al.*, 2016), Malaysia (Rosma, 2010), Japan (Koyanagi *et al.*, 2017), and Taiwan (Hung *et al.*, 2020), and three were from the USA (Lee *et al.*, 2020; Tota *et al.*, 2019) and UK (McCarthy *et al.*, 2020). Sample sizes ranged from 255 to 1,836,888, with cases ranging from 84 to 117,697.

Overall, 15 models were developed from the nine included studies. Some articles reported multiple models (Koyanagi *et al.*, 2017; Lee *et al.*, 2020; Rosma, 2010). Six models focused on oral cancer (Cheung *et al.*, 2021; Hung *et al.*, 2020; Krishna Rao *et al.*, 2016; Lee *et al.*, 2020; Rosma, 2010), three on HNC (Koyanagi *et al.*, 2017; Lee *et al.*, 2020; McCarthy *et al.*, 2020), two on oropharynx (Lee *et al.*, 2020; Tota *et al.*, 2019) and upper aerodigestive tract cancers (Gupta *et al.*, 2017; Koyanagi *et al.*, 2017), and one on hypopharynx (Lee *et al.*, 2020) and larynx (Lee *et al.*, 2020) cancers. The model development methods included Fuzzy regression and Fuzzy Neural Network (Rosma, 2010), Cox Proportional Hazard regression (Cheung *et al.*, 2021), and Multivariable Logistic Regression (Gupta *et al.*, 2017; Hung *et al.*, 2020; Koyanagi *et al.*, 2017; Krishna Rao *et al.*, 2016; Lee *et al.*, 2020; McCarthy *et al.*, 2020; Tota *et al.*, 2019). One article reported the development of a separate model for oesophagus cancer (Koyanagi *et al.*, 2017); we excluded that model as its outcome did not align with our inclusion criteria. Table 2 summarizes model development and assessment techniques in each study. Around 77.8% of the studies (Cheung *et al.*, 2021; Gupta *et al.*, 2017; Koyanagi *et al.*, 2017; Krishna Rao *et al.*, 2016; Lee *et al.*, 2020; McCarthy *et al.*, 2020; Tota *et al.*, 2019) reported missing values, while 22.2% (Hung *et al.*, 2020; Rosma, 2010) did not provide this information. Missing values were handled by imputation techniques in 33.3% of the studies (Cheung *et al.*, 2021; Koyanagi *et al.*, 2017; Tota *et al.*, 2019), while one excluded participants with missing values (Krishna Rao *et al.*, 2016), and another resolved data inconsistencies by communicating with the source dataset investigators (Lee *et al.*, 2020). Only 33.3% (Koyanagi *et al.*, 2017; McCarthy *et al.*, 2020; Tota *et al.*, 2019) of the models were externally validated. All studies used Area Under the Receiver Operating Characteristic Curve (AUC) score to report the model discrimination performance.

Table 3 summarizes the type, outcome, and discriminative performance of each model. Internal and external validation AUC ranged from 0.69 to 0.96 and 0.73 to 0.91, respectively. Gupta et al. (2017) presented the best-performing model (AUC= 95.8 - 95% CI [93.6–97.4]), with positive and negative predictive values of 74.8% and 96.6%, respectively. Regarding calibration, 22.2%

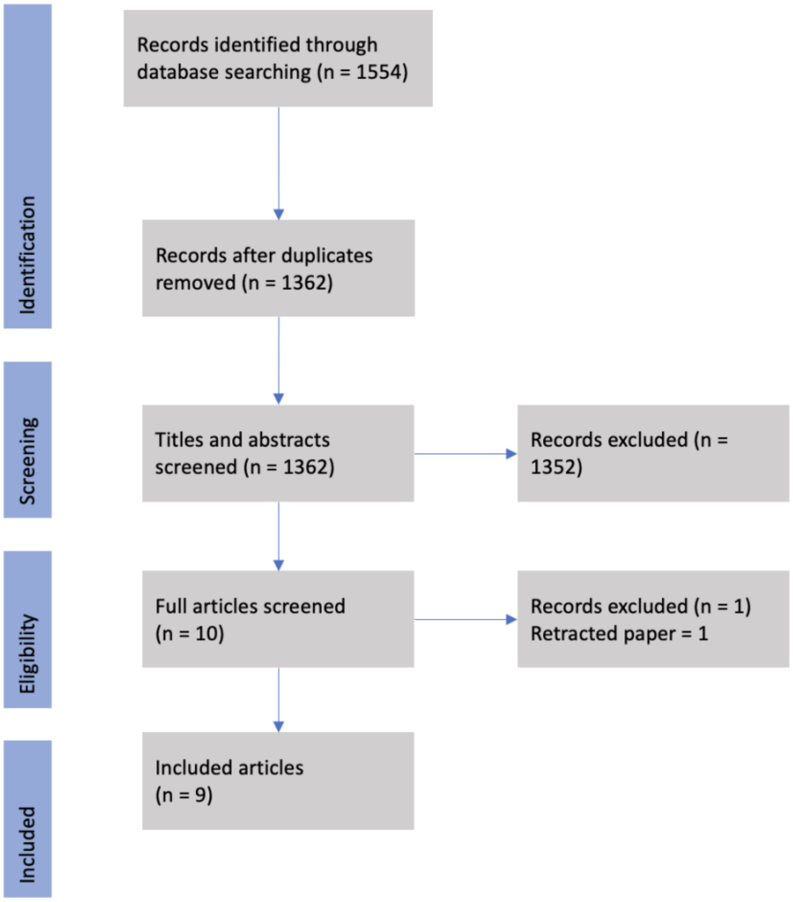
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1554)
Records after duplicates removed (n = 1362)
Titles and abstracts screened (n = 1362)
Records excluded (n = 1352)
Full articles screened (n = 10)
Records excluded (n = 1)
Retracted paper = 1
Included articles (n = 9)

Figure 1*. Flow diagram of the selection process.*

Table 1. General Characteristics of included studies.

First author (year)	*Analysis type*	*Study Design*	*Setting*	*Cases*	*Total*	*Outcome*	*Country of source data*
Cheung (2021)	Cox regression	Cluster-randomized screening trial	Community	395	191,870	Oral cancer incidence	Trivandrum, India
Gupta (2017)	Logistic regression	Case-control	Hospital	240	480	Cancers of lip, oral, oropharynx, hypopharynx, esophagus upper third	Pune, Maharashtra, India
Hung (2020)	Logistic regression	Population based cohort	Community	117,697	1,719,191	Oral cancer incidence	Taiwan
Krishna Rao (2016)	Logistic regression	Case-control	Hospital	180	452	Oral cancer	Karnataka, India
Amy Lee (2020)	Logistic regression	Case-control from registry	Community	7,299	10,301	Cancers of oral, oropharynx, hypopharynx, or larynx	The USA
Tota (2019)	Logistic regression	Case-control from registry	Hospital & Community	241	9,568	Oropharynx cancers	The USA
McCarthy (2020)	Logistic regression	Nested case-control	Community	389	502,177	Head and neck cancer excluding laryngeal cancer	The UK
Koyanagi (2016)	Logistic regression	Case-control	Hospital	1,284	3,198	Cancers of UADT[1], H&N[2], esophageal	Nagoya, Japan
Rosma (2010)	FNN & FR[3]	Case-control	NP[4]	84	171	Oral cancer	Malaysia

[1]Upper aerodigestive tract; [2]Head & Neck; [3]Fuzzy neural network & Fuzzy regression; [4]Not provided

of studies (Koyanagi *et al.*, 2017; Krishna Rao *et al.*, 2016) reported Hosmer-Lemeshow goodness-of-fit (GOF), while 22.2% reported only the calibration score (observed/expected ratio) (Cheung *et al.*, 2021; Tota *et al.*, 2019), and 22.2% demonstrated calibration plots in additional to the score reporting (Lee *et al.*, 2020; McCarthy *et al.*, 2020) (22.2%). Only one study (Koyanagi *et al.*, 2017) provided all three abovementioned calibration measures. Three articles (Gupta *et al.*, 2017; Hung *et al.*, 2020; Rosma, 2010) did not report calibration measurements.

Table 2. Model development characteristics of each study.

First author (year)	*Study type*	*Missing data*	*Missing data management*	*Calibration measurement*	*Outcome frequency adjustment*	*Internal validation*
Cheung (2021)	Development	Yes	Imputation	O/ E[5] (Calibration score)	Yes	Cross validation
Gupta (2017)	Development	Yes	NP[6]	Not provided	NP	Bootstrapping
Hung (2020)	Development	NP	NP	Not provided	Yes (Cohort)	Not provided
Krishna Rao (2016)	Development	Yes	Excluded from analysis	H-L GOF[7] test	Yes	Bootstrapping
Amy Lee (2020)	Development	Yes	Inconsistencies resolved by discussion	Calibration score and plot	Yes	Splitting
Tota (2019)	Development & validation	Yes	Imputation	O/ E (Calibration score)	Yes	Splitting
McCarthy (2020)	Development & validation	Yes	NP	Calibration score and plot	NP	Nothing done
Koyanagi (2016)	Development and validation	Yes	Imputation (coded as dummy variables)	H-L GOF & Calibration plot	Yes	Not provided
Rosma (2010)	Development	NP	NP	Not provided	NP	Splitting

[5]Observed/Expected ratio; [6]Not Provided; [7]Hosmer–Lemeshow goodness of fit

Table 3. Model type, outcome, predictors, and performance of models.

Model	*Model Type*	*Outcome*	*Performance Metrics*
Cheung (2021)	Cox regression	Oral cancer incidence	AUC Overall[8]: 0.84 (0.77–0.90) AUC Ever T&A[9]: 0.75 (0.67–0.83) O/ E[10] Overall: 1.08 (0.81–1.44) O/ E Ever T&A: 1.07 (0.77–1.43)
Gupta (2017)	Multivariable logistic regression	Cancers of lip, oral, oropharynx, hypopharynx, upper third of esophagus	AUC = 95.8 (93.60–97.40) PPV[11]: 74.80% NPV[12]: 96.60%
Hung (2020)	Multivariable logistic regression	Oral cancer incidence	AUC = 0.7306 PPV: 63.90% NPV: 71.10%
Krishna Rao (2016)	Multivariable logistic regression	Oral cancer	AUC = 0.869 PPV: 77.30% NPV: 83.00%
Amy Lee (2020)	Multivariable logistic regression	An invasive tumor of oral cavity, oropharynx, hypopharynx, or larynx	AUC: 0.70
Tota (2019)	Multivariable logistic regression	Oropharynx cancers	Internal[13]: AUC: 0.94 (0.92-0.97) O/ E: 1.05 (0.67-1.44) External[14]: AUC: 0.87 (0.84-0.90) O/ E: 0.91 (0.57-1.25)
McCarthy (2020)	Multivariable logistic regression	Head and neck cancer	AUC: 0.69 (0.66-0.71) Calibration Slope (external): 0.83
Koyanagi (2016)	Conditional logistic regression	Cancers of upper aerodigestive tract, head and neck, esophagus	AUC Internal: 0.59 AUC External: 0.54
Rosma (2010)	Fuzzy neural network & Fuzzy regression	Oral cancer	AUC Fuzzy neural network: 0.804 AUC Fuzzy regression: 0.799 AUC Clinicians' predictions: 0.631

[8]Area under the curve related to the internal validation on overall population; [9]Area under the curve related to the internal validation on ever tobacco and/or alcohol users; [10]Observed/Expected ratio; [11]Positive predictive value; [12]Negative predictive value; [13]Internal validation; [14]External validation

Three studies (Lee *et al.*, 2020; Rosma, 2010; Tota *et al.*, 2019) used splitting, one study (Cheung *et al.*, 2021) used cross-validation, and two studies (Gupta *et al.*, 2017; Krishna Rao *et al.*, 2016) used bootstrapping for internal validation. Three studies (Hung *et al.*, 2020; Koyanagi *et al.*, 2017; McCarthy *et al.*, 2020) did not report internal validation.

Supplementary Table II (Available here: https://doi.org/10.5683/SP3/AV7K47) displays the predictors of the models. The most frequently used predictors were sex (91%), age (88.9%), tobacco smoking (77.8%), alcohol consumption (66.7%), tobacco chewing (44.4%), and education (44.4%). Only one study (Tota *et al.*, 2019) considered HPV a predictor, while another study (Gupta *et al.*, 2017) used lifetime alcohol and tobacco consumption as predictors.

Table 2-4 presents the quality appraisal results, showing a high ROB in the "analysis" domain, mainly affecting three studies (Gupta *et al.*, 2017; McCarthy *et al.*, 2020; Rosma, 2010). All studies but one (Hung *et al.*, 2020) had high ROB in the "predictors" domain, and another lacked sufficient information for assessment (Rosma, 2010). Most studies had low ROB in the "outcome" domain, except two (Hung *et al.*, 2020; Rosma, 2010). The study by Hung *et al.* (2020) raised significant concerns about its applicability (CAA) within the "Participants" domain. In contrast, the study by Cheung *et al.* (2021) demonstrated high CAA in both the "Predictors" and "Outcome" domains.

ROB assessment details are provided in the supplementary figures I and II (Available here: https://doi.org/10.5683/SP3/AV7K47).

Discussion

Nine papers were identified with HNC prognostic risk prediction models. (Cheung *et al.*, 2021; Gupta *et al.*, 2017; Hung *et al.*, 2020; Koyanagi *et al.*, 2017; Krishna Rao *et al.*, 2016; Lee *et al.*, 2020; McCarthy *et al.*, 2020; Tota *et al.*, 2019). All studies had high ROB, mainly due to analytical issues (e.g., missing values, calibration). According to PROBAST, seven studies (Gupta *et al.*, 2017; Koyanagi *et al.*, 2017; Krishna Rao *et al.*, 2016; Lee *et al.*, 2020; McCarthy *et al.*, 2020; Rosma, 2010; Tota *et al.*, 2019) had low applicability concerns. Only three studies (Koyanagi *et al.*, 2017; McCarthy *et al.*, 2020; Tota *et al.*, 2019) externally validated models for clinical application.

The studies were conducted in several countries, with no specific geographical pattern. Among them, two US-based studies (Lee *et al.*, 2020; Tota *et al.*, 2019) developed five models to predict the overall risk of HNC according to subsite. These models are of significance given the recent sharp rise in HPV-related HNC, particularly oropharyngeal cancer, in the US (Chaturvedi *et al.*, 2011).

Koyanagi et al. (2017) developed three models for predicting the risk of cancer of the oropharynx, esophagus, and HNC overall in a Japanese population. Similarly, three India-based studies developed models for oral cancer (Cheung *et al.*, 2021; Krishna Rao *et al.*, 2016) and HNC (Gupta *et al.*, 2017). While these studies help predict HNC in their specific population, they cannot be used in the whole country; India's population is highly diverse (Xing *et al.*, 2010), and thus, the baseline risk should be assessed and adjusted before implementing the models on a different population in that country.

Despite the prevalence of HNC in European and Latin American countries (Winn *et al.*, 2015), only one study from these regions existed (McCarthy *et al.*, 2020), specifically from the UK. There are no prediction models available for the Canadian and Australian populations, despite the diagnosis of 7,400 and 5,104 new cases of HNC in these countries in 2021 (Lee, 2022; Australia, 2022), respectively.

The source of data is vital in risk prediction modelling. The best dataset comes from longitudinal investigations specifically designed for the modelling (Moons *et al.*, 2019). However, these studies are expensive, so the routine practice is to use data from existing cohorts or case-control studies (Lewallen and Courtright, 1998). Nonetheless, secondary data poses challenges such as inconsistency and requiring quality checks before modelling. Also, data from case-control studies need outcome frequency adjustment (Moons *et al.*, 2019). These two challenges must be considered to avoid the risk of biased estimations. Among the seven studies (Gupta *et al.*, 2017; Koyanagi *et al.*, 2017; Krishna Rao *et al.*, 2016; Lee *et al.*, 2020; McCarthy *et al.*, 2020; Rosma, 2010; Tota *et al.*, 2019) using such data, three did not address these

Table 4. PROBAST results.

Author (year)	*ROB*				*Applicability*			*Overall*	
	Participants	*Predictors*	*Outcome*	*Analysis*	*Participants*	*Predictors*	*Outcome*	*ROB*	*Applicability*
Cheung (2021)	+	-	+	-	+	-	-	-	-
Amy Lee (2020)	+	-	+	-	+	+	+	-	+
Hung (2020)	+	+	?	-	-	+	-	-	-
McCarthy (2020)	-	-	+	-	+	+	+	-	+
Rosma (2010)	-	-	?	-	+	+	+	-	+
Tota (2019)	+	-	+	+	+	+	+	-	+
Gupta (2017)	-	-	+	-	+	+	+	-	+
Koyanagi (2016)	+	-	+	-	+	+	+	-	+
Krishna (2016)	+	-	+	-	+	+	+	-	+

challenges (Gupta *et al.*, 2017; McCarthy *et al.*, 2020; Rosma, 2010), thus were classified as high ROB in the "Participants" domain.

Most studies in this review included age, sex, tobacco and alcohol consumption, HPV infection, socioeconomic position, and dietary habits as predictors. Some studies used area-specific risk factors such as Mishri (Gupta *et al.*, 2017), bidi smoking (Gupta *et al.*, 2017), or betel chewing (Hung *et al.*, 2020; Krishna Rao *et al.*, 2016). Incorporating area-specific predictors affects the model's applicability. E.g., models including Mishri do not apply to non–consumer populations. HPV infection is a significant risk factor for a subset of HNC, especially in Western countries (Sabatini and Chiocca, 2020). However, laboratory tests for HPV detection may not always be available. Behavioural factors (e.g., sex behaviour) are a proxy for HPV infection measurement (Tota *et al.*, 2019), which could be used in primary care settings.

Regarding the predictor assessment, the assessor should be blinded to a participant's outcome status (Moons *et al.*, 2019). The studies that used case-control data, in which data collectors are aware of outcomes, led to high ROB in the "predictors" domain assessment.

We identified high ROB and CAA in Cheung et al. (2021) as one of its predictors was not replicable. They used a cluster-randomized controlled screening trial to create the dataset and included the "Screening arm" as a predictor making the model inapplicable to other settings.

Suboptimal outcome ascertainment may lead to misclassification, causing biased performance measurement. Two studies (Hung *et al.*, 2020; Rosma, 2010) had high ROB in the "Outcome" domain due to a lack of detail on outcome assessment.

Most studies had issues with analytical strategies. The following guidelines are suggested to ensure a low ROB in the "analysis" domain (Moons *et al.*, 2019; Steyerberg, 2019):

First, a sufficient sample is needed to ensure enough events per variable (EPV). In one study (Rosma, 2010), 84 case participants used the Fuzzy Neural Network technique to develop the model. We assessed high ROB in this study's "Analysis" domain because machine learning-based models require at least 200 EPV to avoid overfitting (Steyerberg, 2019). The findings of studies with small sample sizes are not applicable in clinical settings.

Second, coding continuous variables as categorized variables causes information loss (Moons *et al.*, 2019; Steyerberg, 2019), although for clinical interpretability, widely accepted cut points can be used to mitigate bias (Moons *et al.*, 2019). Six studies (Hung *et al.*, 2020; Koyanagi *et al.*, 2017; Lee *et al.*, 2020; Rosma, 2010) categorized the "Age" variable, and two studies entirely omitted it (Gupta *et al.*, 2017; Krishna Rao *et al.*, 2016), thus had high ROB in the "analysis" domain.

Third, properly managing missing values is crucial to avoid biased model estimation. One study excluded participants with missing data (Krishna Rao *et al.*, 2016), and four did not provide information in this regard (Gupta *et al.*, 2017; Hung *et al.*, 2020; McCarthy *et al.*, 2020; Rosma, 2010); thus, we assessed them as having high ROB.

Fourth, optimal predictor selection is achieved through nonstatistical methods (literature-based importance or clinical applicability). However, when statistical methods are employed, internally validating the model to mitigate overfitting risks becomes essential. One study (Cheung *et al.*, 2021) used Akaike Information Criterion to find the linear relationship between the outcome and possible predictors. However, the authors conducted cross-validation to assess model optimism, resulting in a low ROB assessment. Conversely, one study (McCarthy *et al.*, 2020) used univariate analysis for predictor selection but did not provide information on optimism checks, leading to a high ROB assessment.

Fifth, the model's accuracy hinges on representing the actual risk of the outcome in the target population. Case-control design can increase EPV but may lead to biased estimations by hiding the true case fraction in the target population. To address this, adjustment for sampling fraction is needed to ensure risk estimations reflect absolute outcome probabilities. Only three of the seven case-control studies reviewed, (Koyanagi *et al.*, 2017; Lee *et al.*, 2020; Tota *et al.*, 2019) reported adjustment for sampling fraction, resulting in a low ROB assessment.

Finally, a predictive model's performance must be assessed appropriately. Although AUC indicates the discriminative ability, we must measure the distance between predicted and actual outcomes, using methods like R2, Brier score, and Hosmer-Lemeshow test to truly judge a model's performance. The positive predictive value (PPV), negative predictive value (NPV), and accuracy must also be measured when implementing the model (e.g., external validation). All studies reported AUC, but none reported other measurements of GOF. Also, none of the studies that externally validated the model (Koyanagi *et al.*, 2017; McCarthy *et al.*, 2020; Tota *et al.*, 2019) reported PPV and NPV and accuracy.

Sixth, a calibration report is essential as it assesses the agreement between predicted probability and observed risk. Only Cheung et al. (2021) adequately considered calibration with time-to-event outcomes. Four articles (Gupta *et al.*, 2017; Hung *et al.*, 2020; Krishna Rao *et al.*, 2016; Rosma, 2010) either did not report or inadequately reported calibration measurements.

Last, developed models must be internally validated on their source dataset to avoid overfitting and overestimating risk. Common internal validation methods include splitting, cross-validation, and bootstrapping. Splitting is inefficient as it reduces the sample size for model derivation, leading to imbalanced outcomes and less reliable performance assessment (Steyerberg, 2019). Three studies (Lee *et al.*, 2020; Rosma, 2010; Tota *et al.*, 2019) used splitting. However, cross-validation could be an alternative method for those studies as it provides more robust internal validation. Three papers (Hung *et al.*, 2020; Koyanagi *et al.*, 2017; McCarthy *et al.*, 2020) lacked information on internal validation. One of them (Hung *et al.*, 2020) (Hung *et al.*, 2020) used a large dataset with sufficient EPV (117,697 cases and 1,719,191 controls) and had low ROB in internal validation, indicating a model less prone to overfitting. Reporting internal validation would have added value to the study. The reviewed studies lacked comprehensive reporting of final model components. Only Total et al. (2019) reported all the necessary components of the model.

The primary goal of a prognostic risk prediction model study is to develop a tool for new settings. To achieve this, models must undergo external validation on different populations. Neglecting this process leads to biased estimation and renders the model unusable in new clinical settings due to unclear performance (Moons *et al.*, 2019; Steyerberg, 2019).

The splitting technique is an internal validation procedure that uses the same sources of data to test the models. It should not be misinterpreted as external validation that involves using datasets from the same population at different times (Steyerberg, 2019). While Koyanagi et al. (2017), McCarthy et al. (2020), and Tota et al. (2019) followed standard methods for external validation, Lee et al. (2020) employed splitting. The authors labelled as «validating the model», but it cannot be considered external validation.

Most studies (91%) (Cheung *et al.*, 2021; Gupta *et al.*, 2017; Hung *et al.*, 2020; Koyanagi *et al.*, 2017; Krishna Rao *et al.*, 2016; Lee *et al.*, 2020; McCarthy *et al.*, 2020; Tota *et al.*, 2019) used statistical techniques to develop a model, standard multivariable logistic regression being the most frequent. However, this approach assumes linearity between exposure and outcome, which is rarely true in real-world scenarios. Additionally, it cannot estimate the effects of exposure time on the outcome. Cox proportional hazard models could be alternatives to help predict disease incidence. Cheung et al. (2021) used this approach.

Big data in medicine enables Artificial Intelligence (AI) for accurate modelling. AI predicts medical conditions like breast cancer, prostate cancer, and diabetes. Rosma et al. (2010) used AI to develop a prognostic model for oral cancers, and the model showed slightly better performance but lacked credibility due to the small sample. Currently, the AI use in HNC focuses on diagnostic, recurrence, or survival models (Huynh *et al.*, 2021; Kazmierski *et al.*, 2021; Peng *et al.*, 2021; Salmanpour *et al.*, 2022). Future work should employ big data and advanced AI techniques (e.g., Bayesian neural network, deep learning, decision tree) for generalizable risk prediction in clinical settings.

High ROB was found in the analysis of all studies. TRIPOD (Collins *et al.*, 2015) was published in 2015. Although the studies were published after 2015, they all had a high ROB in the analysis, mostly due to the lack of proper reporting. Considering that the PROBAST (Moons *et al.*, 2019; Wolff *et al.*, 2019) was published in 2019, there is a need for communicating essential reporting checklists and in prognostic HNC prediction modelling.

We comprehensively evaluated published papers reporting prognostic risk prediction models for HNC and subsites, shedding light on the current status and providing insights for future research. This is the first study in the field that used PROBAST for quality assessment and organized the study using the TRIPOD reporting checklist. Our study is limited by not defining a specific outcome in the research question. We included all types of cancers in the upper aerodigestive tract due to the limited number of modelling papers for each HNC subsite. However, this approach overlooks variations in tumour ethology and behaviour (Morris *et al.*, 2011; Smith *et al.*, 2012). We also didn't report the distribution of predictors due to the limited information the studies provided.

The reproducibility of models is vital for researchers and clinicians. Most papers do not provide enough information to ensure accurate reproducibility. Future studies should fully report the modelling process and share data analysis codes to build on previously developed models to produce reproducible country-specific prognostic HNC risk prediction models.

The ultimate goal of prognostic risk prediction modelling is to develop a tool for identifying high-risk individuals in primary care settings, who can be warned about their high-risk behaviours. These models can also serve as encouraging tools, helping high-risk individuals track risk changes over time when modifying high-risk behaviours like reducing smoking or alcohol consumption.

Conclusion

Many prognostic modelling studies fail to provide sufficient information to judge their models' performance. HNC prognostic risk prediction still needs a well-developed and well-performed model to help clinicians in critical dilemmas. Risk prediction models are complementary tools, and their estimates should not be considered the only means for clinical decision-making. Prognostic risk prediction models are generalizable and applicable only to the source population. Therefore, a model derived from the data related to one specific region in a country (e.g., province or state) does not apply to the whole population of that country. As a result, there is always a need for a well-developed and updated model for each geographical area and population of interest.

Funding statement

The authors received no funds to conduct this project.

References

CAB Abstracts. *CABI.org*. From https://www.cabi.org/publishing-products/cab-abstracts/.

Document search - Web of Science Core Collection*.* From https://www.webofscience.com/wos/woscc/basic-search.

Embase*.* From https://www.wolterskluwer.com/en/solutions/ovid/embase-903.

Scopus preview*.* From https://www.scopus.com/.

Argirion, I., Zarins, K.R., Defever, K., Suwanrungruang, K., Chang, J.T., Pongnikorn, D., Chitapanarux, I., Sriplung, H., Vatanasapt, P. and Rozek, L.S. (2019): Temporal changes in head and neck cancer incidence in Thailand suggest changing oropharyngeal epidemiology in the region. *Journal of Global Oncology* **5**, 1-11.

Arksey, H. and O'Malley, L. (2005): Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* **8**, 19-32.

Australia, C. (2022): *Head and neck cancer in Australia statistics*. From https://www.canceraustralia.gov.au/cancer-types/head-and-neck-cancer/statistics.

Carvalho, A.L., Nishimoto, I.N., Califano, J.A. and Kowalski, L.P. (2005): Trends in incidence and prognosis for head and neck cancer in the United States: a site-specific analysis of the SEER database. *International Journal of Cancer* **114**, 806-816.

Chatterjee, N., Shi, J. and García-Closas, M. (2016): Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* **17**, 392-406.

Chaturvedi, A.K., Engels, E.A., Pfeiffer, R.M., Hernandez, B.Y., Xiao, W., Kim, E., Jiang, B., Goodman, M.T., Sibug-Saber, M., Cozen, W., Liu, L., Lynch, C.F., Wentzensen, N., Jordan, R.C., Altekruse, S., Anderson, W.F., Rosenberg, P.S. and Gillison, M.L. (2011): Human papillomavirus and rising oropharyngeal cancer incidence in the United States. *Journal of Clinical Oncology* **29**, 4294-4301.

Chen, L. (2020): Overview of clinical prediction models. *Annals of Translational Medicine* **8**, 71.

Cheung, L.C., Ramadas, K., Muwonge, R., Katki, H.A., Thomas, G., Graubard, B.I., Basu, P., Sankaranarayanan, R., Somanathan, T. and Chaturvedi, A.K. (2021): Risk-Based Selection of Individuals for Oral Cancer Screening. *Journal of Clinical Oncology* **39**, 663-674.

Cohen, J. (1960): A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* **20**, 37-46.

Collins, G.S., Reitsma, J.B., Altman, D.G. and Moons, K.G.M. (2015): Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *European Urology* **67**, 1142-1151.

Curado, M.P. and Hashibe, M. (2009): Recent changes in the epidemiology of head and neck cancer. *Current Opinion in Oncology* **21**, 194-200.

Dhull, A.K., Atri, R., Dhankhar, R., Chauhan, A.K. and Kaushal, V. (2018): Major Risk Factors in Head and Neck Cancer: A Retrospective Analysis of 12-Year Experiences. *World Journal of Oncology* **9**, 80-84.

Diana, G. and Corica, C. (2021): Human Papilloma Virus vaccine and prevention of head and neck cancer, what is the current evidence? *Oral Oncology* **115**, 105168.

Domchek, S.M., Eisen, A., Calzone, K., Stopfer, J., Blackwood, A. and Weber, B.L. (2003): Application of breast cancer risk prediction models in clinical practice. *Journal of Clinical Oncology* **21**, 593-601.

Gotschall, T. (2021): EndNote 20 desktop version. *J Med Libr Assoc* **109**, 520-522.

Guo, L.-W., Lyu, Z.-Y., Meng, Q.-C., Zheng, L.-Y., Chen, Q., Liu, Y., Xu, H.-F., Kang, R.-H., Zhang, L.-Y., Cao, X.-Q., Liu, S.-Z., Sun, X.-B., Zhang, J.-G. and Zhang, S.-K. (2022): A risk prediction model for selecting high-risk population for computed tomography lung cancer screening in China. *Lung Cancer* **163**, 27-34.

Gupta, B., Kumar, N. and Johnson, N.W. (2017): A risk factor-based model for upper aerodigestive tract cancers in India: predicting and validating the receiver operating characteristic curve. *Journal of Oral Pathology and Medicine* **46**, 465-471.

Habbous, S., Chu, K.P., Lau, H., Schorr, M., Belayneh, M., Ha, M.N., Murray, S., O'Sullivan, B., Huang, S.H., Snow, S., Parliament, M., Hao, D., Cheung, W.Y., Xu, W. and Liu, G. (2017): Human papillomavirus in oropharyngeal cancer in Canada: analysis of 5 comprehensive cancer centres using multiple imputation. *Canadian Medical Association Journal* **189**, E1030-e1040.

Hashim, D., Genden, E., Posner, M., Hashibe, M. and Boffetta, P. (2019): Head and neck cancer prevention: from primary prevention to impact of clinicians on reducing burden. *Annals of Oncology* **30**, 744-756.

Hendriksen, J.M., Geersing, G.-J., Moons, K.G. and de Groot, J.A. (2013): Diagnostic and prognostic prediction models. *Journal of Thrombosis and Haemostasis* **11**, 129-141.

Hung, L.C., Kung, P.T., Lung, C.H., Tsai, M.H., Liu, S.A., Chiu, L.T., Huang, K.H. and Tsai, W.C. (2020): Assessment of the Risk of Oral Cancer Incidence in A High-Risk Population and Establishment of A Predictive Model for Oral Cancer Incidence Using A Population-Based Cohort in Taiwan. *International Journal of Environmental Respiratory Public Health* **17**.

Huynh, B.-N., Ren, J., Groendahl, A.R., Tomic, O., Korreman, S.S. and Futsaether, C.M. (2021). Comparing deep learning and conventional machine learning for outcome prediction of head and neck cancer in PET/CT. *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, Springer**:** 318-326.

Johnson-Obaseki, S., McDonald, J.T., Corsten, M. and Rourke, R. (2012): Head and neck cancer in Canada: trends 1992 to 2007. *Otolaryngology--Head and Neck Surgery* **147**, 74-78.

Joseph, A.W. and D'Souza, G. (2012): Epidemiology of human papillomavirus-related head and neck cancer. *Otolaryngology Clinics of North America* **45**, 739-764.

Kazmierski, M., Welch, M., Kim, S., McIntosh, C., Head, P.M., Group, N.C., Rey-McIntyre, K., Huang, S.H., Patel, T. and Tadic, T. (2021): A Machine Learning Challenge for Prognostic Modelling in Head and Neck Cancer Using Multi-modal Data. *arXiv preprint arXiv:2101.11935*.

Kligerman, M.P., Sethi, R.K.V., Kozin, E.D., Gray, S.T. and Shrime, M.G. (2019): Morbidity and mortality among patients with head and neck cancer in the emergency department: A national perspective. *Head and Neck* **41**, 1007-1015.

Koyanagi, Y.N., Ito, H., Oze, I., Hosono, S., Tanaka, H., Abe, T., Shimizu, Y., Hasegawa, Y. and Matsuo, K. (2017): Development of a prediction model and estimation of cumulative risk for upper aerodigestive tract cancer on the basis of the aldehyde dehydrogenase 2 genotype and alcohol consumption in a Japanese population. *European Journal of Cancer Prevention* **26**, 38-47.

Krishna Rao, S., Mejia, G.C., Logan, R.M., Kulkarni, M., Kamath, V., Fernandes, D.J., Ray, S. and Roberts-Thomson, K. (2016): A screening model for oral cancer using risk scores: development and validation. *Community Dentistry Oral Epidemiology* **44**, 76-84.

Lee, S. (2022): Cancer Statistics. From https://cancer.ca/en/research/cancer-statistics.

Lee, Y.A., Al-Temimi, M., Ying, J., Muscat, J., Olshan, A.F., Zevallos, J.P., Winn, D.M., Li, G., Sturgis, E.M., Morgenstern, H., Zhang, Z.F., Smith, E., Kelsey, K., McClean, M., Vaughan, T.L., Lazarus, P., Chen, C., Schwartz, S.M., Gillison, M., Schantz, S., Yu, G.P., D'Souza, G., Gross, N., Monroe, M., Kim, J., Boffetta, P. and Hashibe, M. (2020): Risk Prediction Models for Head and Neck Cancer in the US Population From the INHANCE Consortium. *American Journal of Epidemiology* **189**, 330-342.

Lewallen, S. and Courtright, P. (1998): Epidemiology in practice: case-control studies. *Community Eye Health* **11**, 57-58.

Marur, S. and Forastiere, A.A. (2008). Head and neck cancer: changing epidemiology, diagnosis, and treatment. *Mayo Clinic Proceedings*, Elsevier. **83:** 489-501.

McCarthy, C.E., Bonnet, L.J., Marcus, M.W. and Field, J.K. (2020): Development and validation of a multivariable risk prediction model for head and neck cancer using the UK Biobank. *International Journal of Oncology* **57**, 1192-1202.

Moons, K.G.M., Wolff, R.F., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J. and Mallett, S. (2019): PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med* **170**, W1-w33.

Morris, L.G., Sikora, A.G., Patel, S.G., Hayes, R.B. and Ganly, I. (2011): Second primary cancers after an index head and neck cancer: subsite-specific trends in the era of human papillomavirus-associated oropharyngeal cancer. *Journal of Clinical Oncology* **29**, 739-746.

Morris, M., Boruff, J.T. and Gore, G.C. (2016): Scoping reviews: establishing the role of the librarian. *Journal of the Medical Libraries Association* **104**, 346-354.

Ouzzani, M., Hammady, H., Fedorowicz, Z. and Elmagarmid, A. (2016): Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews* **5**, 210.

Peng, Z., Wang, Y., Wang, Y., Jiang, S., Fan, R., Zhang, H. and Jiang, W. (2021): Application of radiomics and machine learning in head and neck cancers. *International Journal of Biological Sciences* **17**, 475.

Peters, M.D.J., Marnie, C., Tricco, A.C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C.M. and Khalil, H. (2020): Updated methodological guidance for the conduct of scoping reviews. *JBI Evidence Syntheses* **18**, 2119-2126.

Pulte, D. and Brenner, H. (2010): Changes in survival in head and neck cancers in the late 20th and early 21st century: a period analysis. *Oncologist* **15**, 994-1001.

Rosma, M.D. (2010). The Use of Artificial Intelligence to Identify People at Risk of Oral Cancer : *Empirical Evidence in Malaysian University*.

Sabatini, M.E. and Chiocca, S. (2020): Human papillomavirus as a driver of head and neck cancers. *British Journal of Cancer* **122**, 306-314.

Salmanpour, M.R., Hosseinzadeh, M., Modiri, E., Akbari, A., Hajianfar, G., Askari, D., Fatan, M., Maghsudi, M., Ghaffari, H. and Rezaei, M. (2022). Advanced survival prediction in head and neck cancer using hybrid machine learning systems and radiomics features. *Medical Imaging 2022: Biomedical Applications in Molecular, Structural, and Functional Imaging* SPIE. **12036:** 314-321.

Shipe, M.E., Deppen, S.A., Farjah, F. and Grogan, E.L. (2019): Developing prediction models for clinical use using logistic regression: an overview. *Journal of Thoracic Disease* **11**, S574-s584.

Silveira, A., Monteiro, E. and Sequeira, T. (2018): Head and Neck Cancer: Improving Patient-Reported Outcome Measures for Clinical Practice. *Curr Treat Options Oncol* **19**, 59.

Smith, C.D.L., McMahon, A.D., Ross, A., Inman, G.J. and Conway, D.I. (2022): Risk prediction models for head and neck cancer: A rapid review. *Laryngoscope Investigative Otolaryngology* **7**, 1893-1908.

Smith, E.M., Rubenstein, L.M., Haugen, T.H., Pawlita, M. and Turek, L.P. (2012): Complex etiology underlies risk and survival in head and neck cancer human papillomavirus, tobacco, and alcohol: a case for multifactor disease. *Journal of Oncology* **2012**, 571862.

Steyerberg, E.W. (2019): *Clinical prediction models : a practical approach to development, validation, and updating*. Springer.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F. (2021): Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. A Cancer Journal for Clinicians* **71**, 209-249.

Tammemaegi, M.C. (2015): Application of risk prediction models to lung cancer screening: a review. *Journal of Thoracic Imaging* **30**, 88-100.

Tammemägi, M.C., Ruparel, M., Tremblay, A., Myers, R., Mayo, J., Yee, J., Atkar-Khattra, S., Yuan, R., Cressman, S., English, J., Bedard, E., MacEachern, P., Burrowes, P., Quaife, S.L., Marshall, H., Yang, I., Bowman, R., Passmore, L., McWilliams, A., Brims, F., Lim, K.P., Mo, L., Melsom, S., Saffar, B., Teh, M., Sheehan, R., Kuok, Y., Manser, R., Irving, L., Steinfort, D., McCusker, M., Pascoe, D., Fogarty, P., Stone, E., Lam, D.C.L., Ng, M.Y., Vardhanabhuti, V., Berg, C.D., Hung, R.J., Janes, S.M., Fong, K. and Lam, S. (2022): USPSTF2013 versus PLCOm2012 lung cancer screening eligibility criteria (International Lung Screening Trial): interim analysis of a prospective cohort study. *Lancet Oncology* **23**, 138-148.

Tiwana, M.S., Wu, J., Hay, J., Wong, F., Cheung, W. and Olson, R.A. (2014): 25 year survival outcomes for squamous cell carcinomas of the head and neck: population-based outcomes from a Canadian province. *Oral Oncology* **50**, 651-656.

Toporcov, T.N., Znaor, A., Zhang, Z.-F., Yu, G.-P., Winn, D.M., Wei, Q., Vilensky, M., Vaughan, T., Thomson, P. and Talamini, R. (2015): Risk factors for head and neck cancer in young adults: a pooled analysis in the INHANCE consortium. *International Journal of Epidemiology* **44**, 169-185.

Tota, J.E., Gillison, M.L., Katki, H.A., Kahle, L., Pickard, R.K., Xiao, W., Jiang, B., Graubard, B.I. and Chaturvedi, A.K. (2019): Development and validation of an individualized risk prediction model for oropharynx cancer in the US population. *Cancer* **125**, 4407-4416.

Westphaln, K.K., Regoeczi, W., Masotya, M., Vazquez-Westphaln, B., Lounsbury, K., McDavid, L., Lee, H., Johnson, J. and Ronis, S.D. (2021): From Arksey and O'Malley and Beyond: Customizations to enhance a team-based, mixed approach to scoping review methodology. *Methods X* **8**, 101375.

Winn, D.M., Lee, Y.C., Hashibe, M. and Boffetta, P. (2015): The INHANCE consortium: toward a better understanding of the causes and mechanisms of head and neck cancer. *Oral Diseases* **21**, 685-693.

Wolff, R.F., Moons, K.G.M., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J. and Mallett, S. (2019): PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine* **170**, 51-58.

Xing, J., Watkins, W.S., Hu, Y., Huff, C.D., Sabo, A., Muzny, D.M., Bamshad, M.J., Gibbs, R.A., Jorde, L.B. and Yu, F. (2010): Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biology* **11**, R113.