Title: A risk prediction model for head and neck cancers incorporating lifestyle factors, HPV serology and genetic markers

Sanjeev Budhathoki[1], Brenda Diergaarde[2], Geoffrey Liu[3,4], Andrew Olshan[5], Andrew Ness[6,7], Tim Waterboer[8], Shama Virani[9], Patricia Basta[10], Noemi Bender[8], Nicole Brenner[8], Tom Dudding[7], Neil Hayes[11], Andrew Hope[12], Shao Hui Huang[13], Katrina Hueniken[3], Beatriz Kanterewicz[14], James D McKay[9], Miranda Pring[7], Steve Thomas[7], Kathy Wisniewski[10], Sera Thomas[1], Yonathan Brhane[1], Antonio Agudo[15, 16], Laia Alemany[17], Areti Lagiou[18], Luigi Barzan[19], Cristina Canova[20], David I. Conway[21], Claire M. Healy[22], Ivana Holcatova[23], Pagona Lagiou[24], Gary J. Macfarlane[25], Tatiana V. Macfarlane[25], Jerry Polesel[19], Lorenzo Richiardi[26], Max Robinson[27], Ariana Znaor[28], Paul Brennan[9] and Rayjean J. Hung[1,4]

Affiliations:

[1]Prosserman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Canada

[2]Graduate School of Public Health, University of Pittsburgh, and UPMC Hillman Cancer Center, Pittsburgh, USA

[3]Department of Medical Oncology, Princess Margaret Cancer Centre, University of Toronto, Toronto, Canada

[4]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

[5]University of North Carolina Lineberger Cancer Center, North Carolina, USA

[6]NIHR Bristol Biomedical Research Centre, University of Bristol and Weston NHS Foundation Trust and University of Bristol, UK and Bristol Dental School, University of Bristol, Lower Maudlin St, Bristol, UK

[7]Bristol Dental School, University of Bristol, Lower Maudlin St, Bristol, UK

[8]Infections and Cancer Epidemiology Division, German Cancer Research Center (DKFZ), Heidelberg, Germany

[9]Genetic Epidemiology Group, International Agency for Research on Cancer, Lyon, France

[10]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, USA
[11]Division of Medical Oncology and Center for Cancer Research, University of Tennessee Health Science Center, Memphis, TN, USA
[12]Radiation Oncology, University of Toronto and Radiation Medicine Program, Princess Margaret Cancer Centre, University Health Network, Toronto, Canada
[13]Radiation Oncology, Princess Margaret Cancer Centre, University of Toronto, Toronto, Canada
[14]UPMC Hillman Cancer Center, Pittsburgh, PA, USA
[15]Unit of Nutrition and Cancer, Catalan Institute of Oncology - ICO, L'Hospitalet de Llobregat, Spain.
[16]Nutrition and Cancer Group; Epidemiology, Public Health, Cancer Prevention and Palliative Care Program; Bellvitge Biomedical Research Institute - IDIBELL, L'Hospitalet de Llobregat, Spain.
[17]Catalan Institute of Oncology/IDIBELL, Spain
[18]School of Public Health, University of West Attica, Greece
[19]National Cancer Institute, IRCCS, Italy
[20] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardio-Thoraco-Vascular Sciences and Public Health, University of Padua, Padova, Italy
[21]School of Medicine, Dentistry, and Nursing, University of Glasgow, UK
[22]Trinity College School of Dental Science Dublin, Ireland
[23]Institute of Hygiene and Epidemiology, Czech Republic
[24]School of Medicine, National and Kapodistrian University of Athens, Greece
[25]Epidemiology Group. School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, UK
[26]University of Turin and Reference Centre for Epidemiology and Cancer Prevention in Piemonte, Italy
[27]Centre for Oral Health Research, Newcastle University, UK
[28]International Agency for Research on cancer, France

Correspondence: Rayjean J. Hung, Lunenfeld-Tanenbaum Research Institute, Sinai Health, 60 Murray Street, Toronto, ON M5T3L9, Canada. E-mail: rayjean.hung@lunenfeld.ca

Short title: Risk model for head and neck cancers

Keywords: Head and neck cancer risk, HPV serostatus, polygenic risk score, risk prediction models

Abbreviations: AUC, area under the receiver operating characteristic curves; CI, confidence interval; HPV, human papillomavirus; MFI, median fluorescence intensity; OR, odds ratio; PRS, polygenic risk score.

What's new?

Based on 5 large datasets, this is the first study of an integrated head and neck cancer risk model, including lifestyle risk factors, polygenic risk score, and human papillomavirus serology specifically for oropharyngeal cancer. The models are well-calibrated and showed excellent predictive accuracy. To determine the translational value of these models, we estimated the head and neck cancer absolute risk within the next 5 years using age as the time horizon to determine the optimal time point of actionability. Specifically for oropharyngeal cancer, it showed a distinctive absolute risk trajectory of approximately 3-fold difference for both men and women by risk profiles, with the average risk among human papillomavirus seropositive reaching to 8.1% in men and 2.2% in women at age 60. These risk levels indicate the need of primary prevention or intensive surveillance for the targeted subgroup which is currently lacking.

Abstract

Head and neck cancer is often diagnosed late and prognosis for most head and neck cancer patients remains poor. To aid early detection, we developed a risk prediction model based on demographic and lifestyle risk factors, human papillomavirus (HPV) serological markers, and genetic markers. A total of 10,126 head and neck cancer cases and 5,254 controls from 5 North American and European studies were included. HPV serostatus was determined by antibodies for HPV16 early oncoproteins (E6, E7) and regulatory early proteins (E1, E2, E4). The data were split into a training set (70%) for model development and a hold-out testing set (30%) for model performance evaluation, including discriminative ability and calibration. The risk models including demographic, lifestyle risk factors and polygenic risk score showed a reasonable predictive accuracy for head and neck cancer overall. A risk model that also included HPV serology showed substantially improved predictive accuracy for oropharyngeal cancer (AUC=0.94, 95%CI=0.92-0.95 in men and AUC=0.92, 95%CI=0.88-0.95 in women). The 5-year absolute risk estimates showed distinct trajectories by risk factor profiles. Based on the UK Biobank cohort, the risks of developing oropharyngeal cancer among 60 years old and HPV16 seropositive in the next 5 years ranged from 5.8% to 14.9% with an average of 8.1% for men, 1.3% to 4.4% with an average of 2.2% for women. Absolute risk was generally higher among individuals with heavy smoking, heavy drinking, HPV seropositivity, and those with higher polygenic risk score. These risk models may be helpful for identifying people at high risk of developing head and neck cancer.

Introduction

Head and neck cancer comprises of tumors originating in the oral cavity, hypopharynx, oropharynx, nasopharynx and larynx. In 2020, an estimated 878,348 individuals developed head and neck cancer worldwide, including 65,630 in the United States[1, 2]. The prognosis of head and neck cancer varies by anatomical site and stage at diagnosis. While the 5-year survival rates range from about 50% to 90% for those who are diagnosed at an early stage, patients with head and neck cancer are often diagnosed at an advanced stage, in which case only 20%-40% survive past 5 years[3-5]. In addition, it is well documented that patients with head and neck cancer, particularly advanced stage disease, suffer from significant psychological impact. This is a result of visible disfigurement and disruption of essential functioning due to the disease itself or the treatment [6, 7]. Therefore, early detection and prevention of head and neck cancer is of critical importance.

Tobacco smoking and alcohol consumption are the two well-recognized risk factors for head and neck cancer[5, 8]. Previous pooled analyses have reported more than a two-fold increased risk of head and neck cancer in cigarette smokers and in frequent alcohol drinkers compared to non-users of these substances[9, 10]. Besides tobacco and alcohol use, infection with high-risk HPV types is also an independent risk factor of head and neck cancer. In particular HPV16 is considered a causative agent of head and neck cancer, specifically for cancers of the oropharyngeal region[5, 11-13]. Seropositivity for HPV16 E6 is a highly sensitive and specific marker for HPV-driven oropharyngeal cancer, and blood-based HPV16 E6 antibodies can be found several years before cancer diagnosis[14-17]. In addition, recent genome wide association studies have identified several genetic loci associated with head and neck cancer risk[18, 19].

As a tool to facilitate risk stratification, risk prediction models have been developed using known or potential risk factors for various cancers. Although there were limited previous work reported for risk prediction of head and neck cancer [20-23], none of the existing risk prediction models considered all potential major risk factors, including HPV seropositivity and genetic

susceptibility[20, 21] . In this study, we aimed to develop a prediction model that incorporated demographic, lifestyle, HPV, and identified genetic factors, and to estimate the absolute 5-year risk of developing head and neck cancer, oral cavity cancer and oropharyngeal cancer.

Methods and Materials

Study participants

Five studies from the United States, Canada and Europe were included in this analysis, with a total of 15,380 study participants, including 10,126 head and neck cancer cases and 5254 controls (**Supplementary Figure 1**) from the NIH-funded VOYAGER (Human Papillomavirus, Oral and Oropharyngeal Cancer Genomic Research) program. The five participating studies are Carolina Head and Neck Cancer Epidemiology (CHANCE) and Pittsburgh in the United States, Mount Sinai Hospital-Princess Margaret (MSH-PMH) study in Toronto, Canada, Alcohol-Related Cancers and Genetic susceptibility in Europe (ARCAGE), and Head and neck 5000 (HN5000) in the United Kingdom. The details of these studies have been described previously [24-28]. Briefly, four of the studies are case-control in design and HN5000 is a prospective clinical cohort study with longitudinal follow up of head and neck cancer cases. All cases were patients with squamous cell carcinoma of the head and neck confirmed by pathology reports. Controls were individuals without cancer diagnosis randomly selected from the general population[25, 28], or the visitors of the participating hospitals[24, 27], often frequency-matched to cases in terms of age and sex. All participants were administered a structured questionnaire which assessed information regarding demographic, lifestyle, and medical history. Plasma samples were obtained at the time of diagnosis and prior to start of treatment for cancer cases, and at time of enrollment for controls.

HPV serology assay and genotyping

HPV antibodies were measured in oropharyngeal cancer cases and controls using a bead-based multiplex serology assay [16, 29]. Antigens were affinity-purified, bacterially expressed fusion proteins with N-terminal glutathione S-transferase. We measured antibodies against the early oncoproteins (E6, E7) and regulatory early proteins (E1, E2, E4) for HPV16, and the antibody

values were dichotomized based on predefined median fluorescence intensity (MFI) values[16]. We applied two criteria to determine seropositivity: 1) high antibody levels against HPV 16 E6 alone (>1000 MFI); or 2) seropositivity against three of four HPV16 early proteins (E1: >200 MFI, E2: >679 MFI, E6: >484 MFI and E7: >548 MFI). Participants were considered HPV seropositive if either of these 2 criteria was met[30]. HPV serology was performed at the German Cancer Research Center (DKFZ, Heidelberg, Germany), and laboratory personnel were blinded to the disease status. For controls that were not assayed, we imputed their serostatus by random binomial draw with the overall probability of seropositivity (0.86%) estimated from controls who were assayed[31, 32].

The genetic risk variants included in the model were those previously identified in the genome wide association studies of upper aerodigestive tract cancer risk[18, 19, 33, 34]. A total of 22 variants were included and are summarized in the **Supplementary Table 1**, including 10 for head and neck cancer overall, 5 for oral cavity, and 10 for oropharyngeal cancer. The genotype data for variants were extracted from the head and neck cancer OncoArray dataset previously published[18]. We computed a polygenic risk score (PRS) for head and neck cancer overall and separately for oral cavity cancer and oropharyngeal caner. The PRS was estimated as the sum of the number of risk alleles one carries weighted by the log odds ratio derived from the GWAS studies reported to date except for 8 variants in the HLA region that were identified using HPV-positive oropharyngeal cancer cases[30], where the weights were calculated based on the present study participants.

Exposure variables and cancer endpoints

The demographic and lifestyle factors to be considered in the prediction model were defined *a priori* based on the previous literature. These factors included age, tobacco smoking history (smoking status and pack-years), alcohol consumption history (drinking status and amount of alcohol consumed) and education (postsecondary education as reference). Body mass index was not included in the model, because it was mostly collected at the time of cancer diagnosis and might have been influenced by disease occurrence or progression. In addition to the above predictors, we also included HPV serostatus in the model for oropharyngeal cancer, and PRS in

the model for oral cavity cancer and oropharyngeal cancer. Since a vast majority of the study population (91.4%) self-identified as European descendants, we limited our analysis to those with European ancestry.

All cancer cases were coded according to the International Classification of Disease Volume 10 (ICD-10). In the present analysis, cancer cases were classified as 1) oral cavity cancer: cancers of the lip (C00.3-C00.9, C02.0-C02.3), gum (C03.0, C03.1, C03.9), floor of mouth (C04.0, C04.1, C04.8, C04.9, C05.0) and other and unspecified parts of mouth (C06.0, C06.1, C06.2, C06.8, C06.9); 2) oropharyngeal cancers: cancers of the base of tongue/lingual tonsil (C01.9, C02.4), soft palate (C05.1), uvula (C05.2), palatine tonsil (C09.0, C09.1, C09.8, C09.9) and oropharynx (C10.0, C10.2-C10.9); and 3) other head and neck cancer: cancers of the salivary gland (C07.9-C08.9), nasopharynx (C11.0-C11.9), hypopharynx (C12.9-C13.9), oral cavity-oropharynx-hypopharynx not otherwise specified (C02.8, C02.9, C05.8, C05.9, C14.0, C14.2, C14.8) and larynx (C10.1, C32.0-C32.9). Head and neck cancer included cancers of all the above sites.

Model development and evaluation

Given the substantially different incidence of head and neck cancers by sex, we developed and evaluated the risk model separately for men and women from the outset. For the purpose of model development and evaluation, we randomly divided the data in each study into 70% training set for model development and 30% hold-out testing set for model performance evaluation (**Supplementary Figure 1**). In the training set, we included all statistically significant variables from the univariate logistic regression of the putative risk factors and performed backward stepwise selection to determine the final panel of variables. The linearity was visually inspected by plotting continuous variable against the logit of the outcome and the Box-Tidwell test. Those variables that appear to show a nonlinear relationship were modeled as categorical variables in subsequent analyses. Interactions between variables were evaluated by including product terms of the risk factors in the model. Missing values for lifestyle and demographic variables were imputed using multiple imputation- we created ten imputed datasets by chained equations procedure in which all predictor variables were used to impute missing values. Models were then fitted to each imputed dataset and the results were pooled using Rubin's

rule[35]. Since only two studies had information on family history of head and neck cancer, multiple imputation was not performed for this variable. For variables with multiple measures (such as cigarette and alcohol use status and intensity), we selected a variable based on the Akaike information criterion. The models' ability to discriminate was assessed through Area under the Receiver Operating Characteristic Curves (AUC) in the hold-out testing set. To evaluate the model calibration prospectively on the absolute risk scale, we used the UK Biobank data with longitudinal follow-up (**Supplemental methods**). The model calibration was evaluated by calibration plot comparing the predicted versus the observed probability (defined as empirical proportion of the outcome), and Hosmer-Lemeshow goodness-of-fit test.

Estimation of absolute risk

The five-year absolute risk of developing head and neck cancer was estimated based on Cox proportional hazards model, accounting for age-specific competing hazards of mortality of other causes. The absolute risk within a given time interval was estimated by integrating (i) a model of relative risks, (ii) age-specific incidence of head and neck, oral or oropharyngeal cancer, and (iii) distribution of the risk factors in the population of interest (**Supplementary Methods**). The details of methods have been described in detail previously[36, 37]. The distribution of risk factors was approximated using the UK Biobank population cohort[38, 39]. The age-specific cancer rates and competing hazards for mortality (**Supplementary Table 2**) were obtained from Surveillance, Epidemiology, and End Results (SEER) Program[40] and Centers for Disease Control and Prevention, National Center for Health Statistics database[41] respectively. Since the effect of smoking and alcohol drinking on oropharyngeal cancer may differ by HPV serostatus, we estimated the effect of these risk factors stratified by HPV serostatus, and use the stratum-specific effect estimates for the absolute risk trajectory. A standard non-parametric bootstrap method was used to compute 95% confidence bands of the absolute risk estimates corresponding to the highest risk stratum. Relative risks were estimated from the bootstrap re-samples of the multiple-imputed model building dataset, while age-specific incidence rates, competing mortality rates and the reference dataset were kept constant. All

analyses were performed in R statistical software (version 4.0.3): the *mice* and *psfmi* package for multiple imputation and pooling and *iCARE* package for absolute risk estimation.

Results

The distribution of key characteristics of all study participants are shown in **Table 1**. The study population included more males than females. Cancer cases had higher smoking prevalence and greater pack-years history compared to controls. The average alcohol consumption amount was also higher in cases. As expected, the proportion of participants with HPV seropositivity was much higher among cancer cases, specifically among patients with oropharyngeal cancer. The distributions of risk factor in hypopharynx cancer and larynx cancer showed similar patterns to that of head and neck cancer (**Supplementary Table 3**).

Pack-years and alcohol intensity both showed non-linear association with cancer risk; thus, they were modelled as categorical variables in subsequent analysis. Smoking pack-years was categorised into never, moderate and heavy smokers with the cut-off for the latter two categories being the sex-specific median value of ever smokers among controls (**Supplementary Table 4**). Drinking intensity and PRS were divided into sex-specific tertiles based on the distribution among controls (**Supplementary Table 4**). We did not detect significant interaction between variables and are not included in the final model (**Supplementary Table 5a** and **5b**). Table 2 shows the odds ratios and 95% confidence interval for developing head and neck cancers in the final multivariable model by sex. Overall, in both men and women, smoking, heavy alcohol drinking, lower education, HPV seropositivity, and higher PRS were positively associated with head and neck cancer risk. The association of these factors with oropharyngeal cancer and oral cavity cancer showed similar patterns, albeit the magnitude of the risk estimate was greater for oropharyngeal cancer for smoking and drinking (**Table 2**).

To assess whether the inclusion of a case-only cohort (HN5000) affected our results, we conducted a sensitivity analysis by excluding all HN5000 cases. There was little to no meaningful change of any estimates of the factors included in the model (**Supplementary Table 6**) when including HN5000, thus our primary analysis was based on the full dataset, from which the estimates have higher precision.

The predictive performance of the models in the hold-out testing set based on epidemiological risk factors and the addition of HPV serostatus and PRS is shown in **Table 3** and **Supplementary Figure 2**. In men, the addition of PRS to the model with only epidemiological risk factors improved the discriminative accuracy of the model from AUC of 0.69 to 0.72 (95% CI=0.69-0.75) for head and neck cancer overall, and to 0.73 (95% CI=0.69-0.77) for oral cavity cancer. In women, adding PRS only improved the predictive accuracy for oral cavity cancer, but not for head and neck cancer overall with the resulting AUCs of 0.79 (95% CI=0.74-0.83) and 0.75 (95% CI, 0.71-0.79) respectively. For oropharyngeal cancer, addition of HPV serostatus to the model with only epidemiological risk factor greatly improved the predictive accuracy of the model in both men and women, resulting in the AUCs of 0.92 (95% CI, 0.90-0.94) and 0.91 (95% CI, 0.86-0.94), respectively. Further addition of the PRS marginally improved the predictive accuracy, with AUC of 0.94 (95% CI, 0.92-0.95) in men, and with AUC of 0.92 (95% CI, 0.88-0.95) in women. Assessment of the predictive performance of the models by 10-year age categories showed comparable AUCs in each age strata to the overall AUC for all three cancer types, with small variations, albeit wider confidence intervals. For example, the AUCs of the full model for OPC in women were 0.94 (95%CI, 0.87-0.97), 0.90 (95%CI, 0.81-0.95), 0.91 (95%CI, 0.76-0.97) and 0.91 (95%CI, 0.73-0.98) for age strata of less than 55 years old, 55-64, 65-74 and 75 years and older, respectively.

As a secondary sensitivity analysis, we tested the model performance based on HPV serostatus defined by HPV16 E6 antibody levels (> 1000 MFI) alone to assess the potential loss in predictive accuracy for oropharyngeal cancer. It showed similar AUCs to that of models containing HPV status defined by multiple markers. When HPV seropositivity was defined by HPV16 E6 alone, the AUC for the full model was 0.93 (95%CI=0.90-0.94) in men and 0.89 (95% CI, 0.84-0.93) in women.

Finally, we estimated 5-year absolute risk of head and neck cancer according to risk factor profiles including all aforementioned risk factors included in the final model using the UK Biobank population cohort. The model calibration is shown in **Supplementary Figure 3**. In general, the models are well calibrated based on calibration slope close to 1 and the Hosmer-

Lemeshow test did not indicate deviation for most of the models, except for oropharyngeal cancer in men which has limited sample size and therefore is subject to fluctuations.

Figure 1 shows the absolute risk estimates for overall head and neck cancers. As expected, the absolute risk estimate increased with older age. In general, the risk was low among never users of cigarettes or alcohol in both men and women whereas the risk increased with the heavy use of these substances. The estimated 5-year absolute risk among heavy smokers and heavy drinkers at age 65 varied from 0.64% in the lowest PRS tertile to 1.20% in the highest PRS tertile in men and from 0.23% to 0.30% in women. Since risk profiles are different by anatomical site, we also estimated the 5-year absolute risk separately for oral cavity (**Figure 2**) and oropharyngeal cancer (**Figure 3**). For oral cavity cancer, smoking and drinking accounted for substantial variation in the risk conferred, with those heavy users of both tobacco and alcohol being the highest risk group. In general, the 5-year risk was higher among those with higher PRS in both sexes, but remained low in general (**Figure 2**).

On the other hand, we observed a substantial range of 5-year risk for oropharyngeal cancer and HPV seropositivity status accounted for the majority of the risk variation (**Figure 3**). While the 5-year risk remained very low among those who are HPV-seronegative (<0.1%), the 5-year risk of those who are HPV-seropositives are considerably higher. For example, irrespective of the tobacco and alcohol consumption, the average risk of developing oropharyngeal cancer among HPV seropostives of a 60-year old was 8.1% for men and 2.2% for women (**Figure 3**). In addition, there are differential risk trajectories based on individual's risk profiles. For example, the average 5-year risk for a 60-year old man, HPV-seropositive, lifetime non-drinker and non-smoker was 5.8%, and it increased up to 14.9% for heavy smokers and heavy drinkers, with the other parameters being held constant, albeit wide confidence limits. The corresponding risk estimates for a 60-year old HPV seropositive, lifetime non-drinker and non-smoker woman was 1.3% and in HPV seropositive, heavy smokers and heavy drinkers it was 4.4% (**Figure 3**). For oropharyngeal cancer, due to the very small number of HPV seropositive observations in our control population, we could not estimate the absolute risk by PRS, in conjunction with HPV serostatus.

Discussion

To our knowledge, this is the first study to develop a prediction model for head and neck cancer using HPV serostatus and genetic factors along with known or potential risk factors in European-descent population. The inclusion of HPV serostatus along with epidemiological risk factors improved the model's predictive performance for oropharyngeal cancer. By integrating a US national database of incidence and mortality rates, we observed diverse trajectories by risk factor profiles including HPV serostatus and PRS after accounting for competing risks. Those with HPV seropositive reached high risk level for OPC that could benefit from primary prevention strategy or intensive surveillance, which is currently lacking. These results suggest that risk prediction models can be useful in identifying the population at higher risk of developing head and neck cancer, with the risk varying by anatomical sites and individual risk profiles.

Demographic and lifestyle factors including age, cigarette smoking, alcohol drinking and education were found to be significant predictors of the head and neck cancer risk in our model. The predictive accuracy of our model for oropharyngeal cancer was over 90% when including HPV serostatus, which represents improvements from previous prediction models[20-23]. Given that HPV occurrence is rare for oral cavity cancer, we did not include HPV serostatus as a predictor in the model for oral cavity cancer. However, inclusion of PRS showed modest improved performance for oral cavity cancer, suggesting that combination of multiple risk loci may provide value in oral cancer risk prediction.

HPV16 E6 antibodies are considered to be markers of risk of oropharyngeal cancer. In an analysis using prospectively collected plasma samples from a cohort of European subjects[16], HPV16 E6 seropositivity was associated with a more than 100-fold increase in risk of oropharyngeal cancer[16]. More importantly, this association remained strong based on samples collected more than 10 years before diagnosis[16]. This suggests that HPV16 E6 antibody may have utility as a biomarker for risk stratification of developing oropharyngeal cancer prior to cancer diagnosis. However, the long lead time between HPV seropositivity and cancer diagnosis could pose challenges in screening implementation, with respect to the timing and

frequency of screening and potential psychological burdens due to years of continuous evaluation[13, 42]. On the other hand, the challenges posed by the long lead time of HPV serological markers are not completely distinct from other non-modifiable risk factors such as demographics or genetic susceptibility, which highlights the importance of estimating the absolute risks within a specific time interval using age as the time horizon to determine the optimal time point of actionability, which is the focus of the present study.

In general, screening efficacy depends on pre-cancerous lesions that can be identified with high sensitivity and specificity. Currently, no screening guidelines exist for the early detection of head and neck precancerous lesions or cancers in the general population. For oropharyngeal cancer, while the risk level for the majority of the population is too low to warrant population-based screening, we did observe that the absolute risk trajectory varied greatly by individual's risk factor profiles including smoking, drinking and HPV serostatus. The differentiation of risk trajectory among HPV-seronegatives and HPV-seropositives was predominately depending on the consumption of tobacco and alcohol. We showed that HPV seropositive status led to a high predictive performance, which raises the potential of HPV serology-based test for screening oropharyngeal cancer. In our study, although we used a compound definition of multiple HPV serologic markers, HPV16 E6 seropositivity was the primary driving determinant that defined HPV seropositivity in the majority of participants. Our sensitivity analysis showed that there is limited loss in the predictive accuracy when using HPV16 E6 alone. This suggests that HPV16 E6 antibody is an adequate test to determine the seropositivity, which may help to improve the feasibility of large-scale population testing.

However, the main challenges remain that pre-cancer lesions for oropharyngeal caner have not been identified[42], and given the relatively low incidence of oropharyngeal cancer, the HPV serology-based test would result in low positive predictive value. Both of these factors would limit the balance between psychologic and physical distress related to screening and the potential benefits[13, 43]. Given the low prevalence of HPV16 early protein antibodies in the general population, further studies are required to evaluate the effectiveness of screening modalities in secondary prevention of oropharyngeal cancer, as well as the risk-threshold to maximize the cost-efficiency, which is beyond the scope of the present work. Nonetheless, in

terms of primary prevention, our model may be informative for individuals at high-risk and potentially encouraging behavioral modifications, such as intensive smoking cessation programs.

Regarding oral cancer, the US Preventive Services Task Force concluded that the current evidence was insufficient to assess the balance of benefits and harms of screening for oral cancer in asymptomatic adults[44]. A large trial conducted in India, where participants were randomly assigned to receive visual screening (of the oral cavity by trained healthcare workers every three years for four rounds) *versus* the usual care (control group), reported reduced mortality from oral cancers in the screened group which was mainly observed in tobacco and alcohol users[45]. In another report of a nationwide, population-based screening program for oral cancer in Taiwan, the mortality of oral cancer was reduced by 50% in the screening group compared to the expected oral cancer mortality in the absence of screening[46]. These studies suggest benefit of screening for oral cancer in high-risk groups. However, these studies were conducted in populations with higher incidence of oral cancers and may not be directly generalizable to other populations with different risk profiles. Although other visual adjunctive technologies such as toluidine blue, brush biopsy or fluorescence imaging have been evaluated for oral cancer screening, their effectiveness as a screening tool to reduce oral cancer mortality is not established[47, 48].

Our study has several limitations. First, the study participants represented a population of European ancestry and thus the model may not be generalizable to other ethnicities with different risk factor profiles. Nonetheless, in comparison to the large national survey[49], we found that the risk profiles, mainly cigarette smoking and alcohol drinking in our VOYAGER study is comparable with the large national survey data: 19.3% of the population were current smoker in the survey versus 20.9% in our study, and 78.9% of the population were ever drinkers in the survey population compared to 79.4 in our study. If there was bias introduced due to the source data for risk factor distribution, it would likely to be minimal. On the other hand, the absolute risk was estimated based on UK Biobank population cohort, which has been recognized as a healthier cohort[50]. Therefore, the estimated absolute risks maybe lower than in the general population[50]. Second, even though there was large number of cases for overall

head and neck cancer analysis, the sample size was small for analysis by subsite, particularly with HPV serostatus and genetic data. Cautious interpretation of the extreme high-risk group is needed given the wide confidence bands, in particular for oropharyngeal cancer. Third, only a subset of study participants had information on family history of head and neck cancer in our study and thus this variable was not included in the model. Fourth, our data contributed to the original discovery of the susceptibility loci of head and neck cancers, therefore the PRS effect may be overfitted. Future studies should use independent datasets to reduce the possibility of overfitting in the PRS model. Finally, we were not able to conduct external validation of our model given the limited data availability that include both HPV serology and genetic data available outside of the current participating studies.

In summary, we developed the first absolute risk prediction model for head and neck cancer which incorporated all key aspects including environmental risk factors, HPV serostatus, and genetic risk variants. The model performance was improved compared to previous models based on epidemiologic factors only, and it may be useful for stratifying populations at high risk of developing head and neck cancer. Future validation of these models based on prospective cohorts would be warranted. Nonetheless, the high absolute risk level among those with HPV seropositive highlights the need to consider primary prevention and intensive surveillance for OPC in targeted subgroup.

Funding

This project was funded in part by NIH/NIDCR R01 DE025712 (PB, BD and NH), and the Canada Research Chair from the Canadian Institute of Health Research (RJH).

Acknowledgement

Genotyping of cases and controls was performed at the Center for Inherited Disease Research (CIDR) and funded by NIH/NIDCR 1X01HG007780-0. The MSH-PMH study was supported by Canadian Cancer Society Research Institute and Lusi Wong Programs at the Princess Margaret Hospital Foundation. The University of Pittsburgh head and neck cancer case-control study is

supported by US National Institutes of Health grants P50CA097190 and P30CA047904. The Carolina Head and Neck Cancer Epidemiology (CHANCE) study was supported in part by the National Cancer Institute (R01-CA90731). The Head and Neck 5000 study was a component of independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme (RP-PG-0707-10034). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. Core funding was also provided through awards from Above and Beyond, University Hospitals Bristol and Weston Research Capability Funding and the NIHR Senior Investigator award to Professor Andy Ness. Human papillomavirus (HPV) serology was supported by a Cancer Research UK Programme Grant, the Integrative Cancer Epidemiology Programme (grant number: C18281/A19169). Sanjeev Budhathoki is supported by the Hold'em for Life Oncology Fellowship. Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization. We thank Dr. Wolfgang Ahrens, PhD (University of Bremen, Germany) for his support in ARCAGE study.

Conflict of Interest

Tim Waterboer serves on advisory boards for MSD (Merck) Sharp & Dohme. All other authors report no potential conflicts of interest.

Data Availability Statement

Data sources and handling of the publicly available datasets used in this study are described in the Materials and Methods. Further details and other data that support the findings of this study are available from the corresponding authors upon request.

Author contributions

The work reported in the paper has been performed by the authors, unless clearly specified in the text. R.J. H. conceived the research question and designed the study with S. B. S. B. and Y. B. performed the statistical analysis. B. D., G. L., A. O., A. N., P. Br. and R.J. H. were responsible for funding acquisition and collected the study data. T. W. and N. Br performed the HPV serology assays. S. B. drafted the manuscript with scientific input from R. J. H. All authors contributed to the interpretation of the results and critical revision of the manuscript.

Ethics statement

All participants provided written informed consents, and research protocols of all studies were reviewed and approved by the local institutional review boards of each participating study. This project was approved by the Research Ethics Board at the Sinai Health.

Reference

1. International Agency for Research on Cancer GCO, Cancer Today. Lyon IARC, 2021. Available at: https://gco.iarc.fr/today/home (Accessed March, 2021) 2020.

2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin* 2020;**70**: 7-30.

3. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. *CA Cancer J Clin* 2021;**71**: 7-33.

4. SEER Cancer Stat Facts: Oral Cavity and Pharynx Cancer. National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/statfacts/html/oralcav.html.

5. Thun M, Linet M, Cerhan J, Haiman C, Schottenfeld D. *Cancer Epidemiology and Prevention*, 4th ed.: Oxford University Press, 2017.

6. Valdez JA, Brennan MT. Impact of Oral Cancer on Quality of Life. *Dent Clin North Am* 2018;**62**: 143-54.

7. Cohen EE, LaMonte SJ, Erb NL, Beckman KL, Sadeghi N, Hutcheson KA, Stubblefield MD, Abbott DM, Fisher PS, Stein KD, Lyman GH, Pratt-Chapman ML. American Cancer Society Head and Neck Cancer Survivorship Care Guideline. *CA Cancer J Clin* 2016;**66**: 203-39.

8. Anantharaman D, Muller DC, Lagiou P, Ahrens W, Holcátová I, Merletti F, Kjærheim K, Polesel J, Simonato L, Canova C, Castellsague X, Macfarlane TV, et al. Combined effects of smoking and HPV16 in oropharyngeal cancer. *Int J Epidemiol* 2016;**45**: 752-61.

9. Hashibe M, Brennan P, Benhamou S, Castellsague X, Chen C, Curado MP, Dal Maso L, Daudt AW, Fabianova E, Fernandez L, Wünsch-Filho V, Franceschi S, et al. Alcohol drinking in never users of tobacco, cigarette smoking in never drinkers, and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *J Natl Cancer Inst* 2007;**99**: 777-89.

10. Berthiller J, Straif K, Agudo A, Ahrens W, Bezerra Dos Santos A, Boccia S, Cadoni G, Canova C, Castellsague X, Chen C, Conway D, Curado MP, et al. Low frequency of cigarette smoking and the risk of head and neck cancer in the INHANCE consortium pooled analysis. *Int J Epidemiol* 2016;**45**: 835-45.

11. IARC. Biological agents, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. *IARC Press, World Health Organization* 2012;**Volume 100B**.

12. Gillison ML, Koch WM, Capone RB, Spafford M, Westra WH, Wu L, Zahurak ML, Daniel RW, Viglione M, Symer DE, Shah KV, Sidransky D. Evidence for a Causal Association Between Human Papillomavirus and a Subset of Head and Neck Cancers. *JNCI: Journal of the National Cancer Institute* 2000;**92**: 709-20.

13. Kreimer AR, Shiels MS, Fakhry C, Johansson M, Pawlita M, Brennan P, Hildesheim A, Waterboer T. Screening for human papillomavirus-driven oropharyngeal cancer: Considerations for feasibility and strategies for research. *Cancer* 2018;**124**: 1859-66.

14. Kreimer AR, Ferreiro-Iglesias A, Nygard M, Bender N, Schroeder L, Hildesheim A, Robbins HA, Pawlita M, Langseth H, Schlecht NF, Tinker LF, Agalliu I, et al. Timing of HPV16-E6 antibody seroconversion before OPSCC: findings from the HPVC3 consortium. *Ann Oncol* 2019;**30**: 1335-43.

15. Kreimer AR, Johansson M, Yanik EL, Katki HA, Check DP, Lang Kuhs KA, Willhauck-Fleckenstein M, Holzinger D, Hildesheim A, Pfeiffer R, Williams C, Freedman ND, et al. Kinetics of the Human Papillomavirus Type 16 E6 Antibody Response Prior to Oropharyngeal Cancer. *J Natl Cancer Inst* 2017;**109**.

16. Kreimer AR, Johansson M, Waterboer T, Kaaks R, Chang-Claude J, Drogen D, Tjønneland A, Overvad K, Quirós JR, González CA, Sánchez MJ, Larrañaga N, et al. Evaluation of human papillomavirus antibodies and risk of subsequent head and neck cancer. *J Clin Oncol* 2013;**31**: 2708-15.

17. Hibbert J, Halec G, Baaken D, Waterboer T, Brenner N. Sensitivity and Specificity of Human Papillomavirus (HPV) 16 Early Antigen Serology for HPV-Driven Oropharyngeal Cancer: A Systematic Literature Review and Meta-Analysis. *Cancers* 2021;**13**.

18. Lesseur C, Diergaarde B, Olshan AF, Wünsch-Filho V, Ness AR, Liu G, Lacko M, Eluf-Neto J, Franceschi S, Lagiou P, Macfarlane GJ, Richiardi L, et al. Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nat Genet* 2016;**48**: 1544-50.

19. McKay JD, Truong T, Gaborieau V, Chabrier A, Chuang SC, Byrnes G, Zaridze D, Shangina O, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, et al. A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. *PLoS Genet* 2011;**7**: e1001333.

20. Koyanagi YN, Ito H, Oze I, Hosono S, Tanaka H, Abe T, Shimizu Y, Hasegawa Y, Matsuo K. 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. *Eur J Cancer Prev* 2017;**26**: 38-47.

21. Iwasaki M, Budhathoki S, Yamaji T, Tanaka-Mizuno S, Kuchiba A, Sawada N, Goto A, Shimazu T, Inoue M, Tsugane S, Group JPHC-bPSJS. Inclusion of a gene-environment interaction between alcohol consumption and the aldehyde dehydrogenase 2 genotype in a risk prediction model for upper aerodigestive tract cancer in Japanese men. *Cancer Sci* 2020;**111**: 3835-44.

22. McCarthy CE, Bonnet LJ, Marcus MW, Field JK. Development and validation of a multivariable risk prediction model for head and neck cancer using the UK Biobank. *Int J Oncol* 2020.

23. Lee YA, Al-Temimi M, Ying J, Muscat J, Olshan AF, Zevallos JP, Winn DM, Li G, Sturgis EM, Morgenstern H, Zhang ZF, Smith E, et al. Risk Prediction Models for Head and Neck Cancer in the US Population From the INHANCE Consortium. *Am J Epidemiol* 2020;**189**: 330-42.

24. Macfarlane TV, Macfarlane GJ, Oliver RJ, Benhamou S, Bouchardy C, Ahrens W, Pohlabeln H, Lagiou P, Lagiou A, Castellsague X, Agudo A, Merletti F, et al. The aetiology of upper aerodigestive tract cancers among young adults in Europe: the ARCAGE study. *Cancer Causes Control* 2010;**21**: 2213-21.

25. Bradshaw PT, Siega-Riz AM, Campbell M, Weissler MC, Funkhouser WK, Olshan AF. Associations between dietary patterns and head and neck cancer: the Carolina head and neck cancer epidemiology study. *Am J Epidemiol* 2012;**175**: 1225-33.

26. Beynon RA, Lang S, Schimansky S, Penfold CM, Waylen A, Thomas SJ, Pawlita M, Waterboer T, Martin RM, May M, Ness AR. Tobacco smoking and alcohol drinking at diagnosis of head and neck cancer and all-cause mortality: Results from head and neck 5000, a prospective observational cohort of people with head and neck cancer. *Int J Cancer* 2018;**143**: 1114-27.

27. Troy JD, Grandis JR, Youk AO, Diergaarde B, Romkes M, Weissfeld JL. Childhood passive smoke exposure is associated with adult head and neck cancer. *Cancer Epidemiol* 2013;**37**: 417-23.

28. Thomas S, Carroll JC, Brown MC, Chen Z, Mirshams M, Patel D, Boyd K, Pierre A, Goldstein DP, Giuliani ME, Xu W, Eng L, et al. Nicotine dependence as a risk factor for upper aerodigestive tract (UADT) cancers: A mediation analysis. *PLoS One* 2020;**15**: e0237723.

29. Waterboer T, Sehr P, Michael KM, Franceschi S, Nieland JD, Joos TO, Templin MF, Pawlita M. Multiplex human papillomavirus serology based on in situ-purified glutathione s-transferase fusion proteins. *Clin Chem* 2005;**51**: 1845-53.

30. Ferreiro-Iglesias A, McKay J, Brenner N, Virani S, Lesseur C, Gaborieau V, Ness AR, Hung RJ, Liu G, Diergaarde B, Olshan A, Hayes N, et al. Germline Determinants of Humoral Immune Response To HPV-16 Protect Against Oropharyngeal Cancer. *Nature Communications* 2021;**Accepted/In press - 6 Jul 2021**.

31. Brenner N, Mentzer AJ, Hill M, Almond R, Allen N, Pawlita M, Waterboer T. Characterization of human papillomavirus (HPV) 16 E6 seropositive individuals without HPV-associated malignancies after 10 years of follow-up in the UK Biobank. *EBioMedicine* 2020;**62**: 103123.

32. Lang Kuhs KA, Anantharaman D, Waterboer T, Johansson M, Brennan P, Michel A, Willhauck-Fleckenstein M, Purdue MP, Holcátová I, Ahrens W, Lagiou P, Polesel J, et al. Human Papillomavirus 16 E6 Antibodies in Individuals without Diagnosed Cancer: A Pooled Analysis. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2015;**24**: 683-9.

33. Azad AK, Bairati I, Qiu X, Girgis H, Cheng L, Waggott D, Cheng D, Mirshams M, Ho J, Fortin A, Vigneault E, Huang SH, et al. A genome-wide association study of non-HPV-related head and neck squamous cell carcinoma identifies prognostic genetic sequence variants in the MAP-kinase and hormone pathways. *Cancer Epidemiol* 2016;**42**: 173-80.

34. Buniello A, MacArthur J, Cerezo M, Harris L, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* 2019;**Vol. 47 (Database issue): D1005-D1012**.

35. Rubin DB. Inference and missing data. *Biometrika* 1976;**63**.

36. Gail MH. Estimation and interpretation of models of absolute risk from epidemiologic data, including family-based studies. *Lifetime Data Anal* 2008;**14**: 18-36.

37. Pal Choudhury P, Maas P, Wilcox A, Wheeler W, Brook M, Check D, Garcia-Closas M, Chatterjee N. iCARE: An R package to build, validate and apply absolute risk models. *PLoS One* 2020;**15**: e0228198.

38. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**: e1001779.

39. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**: 203-9.

40. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Data, 9 Registries, Nov 2020 Sub (1975-2018) - Linked To County Attributes - Time Dependent (1990-2018) Income/Rurality, 1969-2019 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2021, based on the November 2020 submission.

41. Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death 1999-2019 on CDC WONDER Online Database, released in 2020. Data are from the Multiple Cause of Death Files, 1999-2019, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/ucd-icd10.html on Jul 19, 2021 12:15:53 PM.

42. Kreimer AR, Chaturvedi AK, Alemany L, Anantharaman D, Bray F, Carrington M, Doorbar J, D'Souza G, Fakhry C, Ferris RL, Gillison M, Neil Hayes D, et al. Summary from an international cancer seminar focused on human papillomavirus (HPV)-positive oropharynx cancer, convened by scientists at IARC and NCI. *Oral oncology* 2020;**108**: 104736.

43. Hashim D, Genden E, Posner M, Hashibe M, Boffetta P. Head and neck cancer prevention: from primary prevention to impact of clinicians on reducing burden. *Ann Oncol* 2019;**30**: 744-56.

44. Moyer VA. Screening for oral cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med* 2014;**160**: 55-60.

45. Sankaranarayanan R, Ramadas K, Thara S, Muwonge R, Thomas G, Anju G, Mathew B. Long term effect of visual screening on oral cancer incidence and mortality in a randomized trial in Kerala, India. *Oral oncology* 2013;**49**: 314-21.

46. Chuang SL, Su WW, Chen SL, Yen AM, Wang CP, Fann JC, Chiu SY, Lee YC, Chiu HM, Chang DC, Jou YY, Wu CY, et al. Population-based screening program for reducing oral cancer mortality in 2,334,299 Taiwanese cigarette smokers and/or betel quid chewers. *Cancer* 2017;**123**: 1597-609.

47. Patton LL, Epstein JB, Kerr AR. Adjunctive techniques for oral cancer examination and lesion diagnosis: a systematic review of the literature. *Journal of the American Dental Association (1939)* 2008;**139**: 896-905; quiz 93-4.

48. Brocklehurst P, Kujan O, O'Malley LA, Ogden G, Shepherd S, Glenny AM. Screening programmes for the early detection and prevention of oral cancer. *The Cochrane database of systematic reviews* 2013: Cd004150.

49. Schiller JS, Lucas JW, Ward BW, Peregoy JA. Summary health statistics for U.S. adults: National Health Interview Survey, 2010. *Vital and health statistics Series 10, Data from the National Health Survey* 2012: 1-207.

50. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R, Allen NE. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017;**186**: 1026-34.

Figure legends

Figure 1. Five-year absolute risk estimates of head and neck cancer stratified by tobacco smoking, alcohol drinking and polygenic risk score for men and women.

The blue, yellow and red lines represent never, moderate and heavy tobacco smokers, respectively. The dashed and solid lines represent never/moderate and heavy alcohol drinkers. For example, yellow solid line represents moderate smokers who drank heavily. The gray zone represents the 95% confidence intervals of the highest risk category. The smoking category (Moderate vs Heavy) cut-off is based on sex-specific medians among ever smokers in the control group. The alcohol drinking categories (Low, Moderate, Heavy) and polygenic risk score (Low, Medium and High) are based on sex-specific tertiles in the control group (Supplementary Table 4).

Figure 2. Five-year absolute risk estimates of oral cavity cancer stratified by smoking, drinking and polygenic risk score for men and women.

The blue, yellow and red lines represent never, moderate and heavy tobacco smokers, respectively. The dashed and solid lines represent never/moderate and heavy alcohol drinkers. For example, yellow solid line represents moderate smokers who drank heavily. The gray zone represents the 95% confidence intervals of the highest risk category. The smoking category (Moderate vs Heavy) cut-off is based on sex-specific medians among ever smokers in the control group. The alcohol drinking categories (Low, Moderate, Heavy) and polygenic risk score (Low, Medium and High) are based on sex-specific tertiles in the control group (Supplementary Table 4).

Figure 3. Absolute risk estimates of oropharyngeal cancer stratified by tobacco smoking, alcohol drinking and human papillomavirus (HPV) serostatus for men and women.

The color of the lines represents different smoking and drinking categories. The solid and dashed line represent HPV seropositive and seronegative, respectively. The dotted line represents the average risk

among HPV seropositive individuals, irrespective of their tobacco and alcohol consumption status. The smoking category (Moderate vs Heavy) cut-off is based on sex-specific medians among ever smokers in the control group. The alcohol drinking categories (Low, Moderate, Heavy) and polygenic risk score (Low, Medium and High) are based on sex-specific tertiles in the control group (Supplementary Table 4).

Table 1. The key characteristics of the study populations

Variables	Categories	Head and neck cancer	Oral cavity cancer	Oropharyngeal cancer	Controls
Total (n)		10126	2431	3727	5254
Study (n)					
	CHANCE	1010	158	277	1114
	ARCAGE	1924	470	411	2043
	PITTSBURGH	847	263	365	811
	TORONTO	1663	400	790	1286
	HN5000	4682	1140	1884	-
Sex, n (%)					
	Men	7750 (76.6)	1572 (64.7)	2976 (79.8)	3471 (66.1)
	Women	2373 (23.4)	859 (35.3)	751 (20.2)	1783 (33.9)
	Missing	3	0	0	0
Age (years), mean (SD)		60.7 (10.9)	61.8 (12.0)	58.7 (9.3)	60 (12.0)
Tobacco Smoking status, n (%)					
	Never	1638 (18.9)	448 (21.5)	780 (24.9)	2135 (40.8)
	Former	3510 (40.5)	744 (35.6)	1346 (43.0)	2009 (38.4)
	Current	3527 (40.7)	895 (42.9)	1006 (32.1)	1094 (20.9)
	Missing	1447	342	592	16
Tobacco Pack-years, median (IQR)		36 (33.0)	34.1 (32.5)	30.0 (32.0)	21.0 (30.5)
Alcohol drinking status, n (%)					
	Never	1797 (20.7)	484 (23.2)	665 (21.0)	1080 (20.6)
	Former	4351 (50.1)	1015 (48.6)	1793 (56.7)	1493 (28.5)
	Current	2531 (29.2)	588 (28.2)	702 (22.2)	2667 (50.9)
	Missing	1443	343	563	14
Drink/week, median (IQR)		20.3 (29.0)	20.5 (29.0)	17.9 (28.0)	7.4 (12.6)
Education, n (%)					
	Postsecondary	2377 (29.8)	523 (27.1)	1043 (35.3)	2585 (52.1)
	High school diploma	2419 (30.4)	622 (32.2)	997 (33.7)	1030 (20.8)
	None/elementary	3168 (39.8)	785 (40.7)	918 (31.0)	1345 (27.1)
	Missing	2163	498	766	294
HPV serostatus, n (%)[a]					
	Total tested			1804	2332
	Negative			660 (36.6)	2312 (99.1)
	Positive			1144 (63.4)	20 (0.9)
Polygenic risk score, median (IQR)	Total genotyped	3901	1339	1823	2962
		0.47 (0.05-0.82)	-0.002 (-0.21 – 0.16)	0.21 (-0.26 – 0.62)	0.24 (-0.26 – 0.66)

[a]HPV serology status is defined based on high HPV16 E6 antibody levels (>1000 median fluorescence intensity, MFI) or seropositivity for three of four HPV16 early proteins (E1: >200 MFI, E2: >679 MFI, E6: >484 MFI and E7: >548 MFI).

Table 2. Odds ratios (ORs) and 95% confidence intervals (CIs) for developing head and neck cancers and key risk factors by sex based on multivariable logistic regression models

Variable	Categories	Head and neck cancer	Oral cavity cancer	Oropharyngeal cancer
		OR (95%CI)*	OR (95%CI)*	OR (95%CI)*
Men				
Smoking status[a]				
	Never	1 (Ref.)	1 (Ref.)	1 (Ref.)
	Moderate	1.22 (1.02-1.44)	1.53 (1.14-2.05)	1.48 (1.00-2.20)
	Heavy	2.52 (2.11-3.01)	3.15 (2.45-4.06)	5.14 (3.54-7.47)
Drinking status[b]				
	Never/Low	1 (Ref.)	1 (Ref.)	1 (Ref.)
	Moderate	0.87 (0.74-1.03)	0.88 (0.69-1.13)	0.83 (0.57-1.20)
	Heavy	1.66 (1.41-1.94)	1.82 (1.46-2.27)	2.27 (1.69-3.04)
Education				
	Postsecondary	1 (Ref.)	1 (Ref.)	1 (Ref.)
	High school diploma	2.26 (1.93-2.65)	2.74 (2.14-3.51)	1.93 (1.40-2.65)
	None/elementary	1.96 (1.65-2.34)	2.71 (2.09-3.51)	2.51 (1.85-3.40)
HPV serostatus				
	Negative			1 (Ref.)
	Positive			385 (218-681)
Polygenic risk score[c]				
	Low (1st tertile)	1 (Ref.)	1 (Ref.)	1 (Ref.)
	Middle (2nd tertile)	1.52 (1.29-1.79)	1.33 (1.05-1.69)	1.14 (0.85-1.55)
	High (3rd tertile)	2.35 (2.01-2.75)	2.16 (1.72-2.71)	1.59 (1.19-2.13)
Women				
Smoking status[a]				
	Never	1 (Ref.)	1 (Ref.)	1 (Ref.)
	Moderate	1.32 (1.01-1.71)	1.37 (0.96-1.95)	2.09 (1.09-4.00)
	Heavy	3.46 (2.75-4.35)	3.23 (2.41-4.32)	6.86 (4.14-11.36)
Drinking status[b]				
	Never/low	1 (Ref.)	1 (Ref.)	1 (Ref.)
	Moderate	0.70 (0.53-0.92)	0.73 (0.51-1.06)	1.12 (0.64-1.97)
	Heavy	1.49 (1.18-1.89)	1.50 (1.11-2.03)	2.64 (1.66-4.18)
Education				
	Postsecondary	1 (Ref.)	1 (Ref.)	1 (Ref.)
	High school diploma	3.17 (2.52-4.00)	3.81 (2.80-5.18)	3.56 (2.17-5.84)
	None/elementary	4.88 (3.77-6.32)	6.44 (4.69-8.84)	5.04 (2.99-8.50)
HPV serostatus				
	Negative			1 (Ref.)
	Positive			237 (103-550)
Polygenic risk score[c]	Low (1st tertile)	1 (Ref.)	1 (Ref.)	1 (Ref.)
	Median (2nd tertile)	1.71 (1.33-2.19)	1.71 (1.24-2.37)	0.90 (0.54-1.52)
	High (2nd tertile)	1.89 (1.48-2.42)	2.07 (1.51-2.84)	1.25 (0.77-2.01)

*The odds ratio estimates are based on all factors included in this table in the multivariable model. [a]The cut-off is based on sex-specific medians among ever smokers in the control group: <24 pack-years (Moderate smoker) or ≥24 pack-years of smoking (Heavy smoker) in men; and <14 pack-years (Moderate smoker) or ≥14 pack-years of smoking (Heavy smoker) in women. [b]The cut-off is based on sex-specific tertiles in the control group; <5.5 drinks/week (Never/low drinker), 5.5 to <14.7 drinks/week (Moderate drinker) or ≥14.7 drinks/week (Heavy drinker) in men; and <2.2 drinks/week (Never/low drinker), 2.2 to <6.9 drinks/week (Moderate drinker) or ≥6.9 drinks/week (Heavy drinker) in women. [c]The

polygenic risk scores are computed for oral cavity and oropharyngeal cancer separately based on the loci reported for these tumor types. Loci reported for head and neck cancer or their anatomical subsites are included in the PRS for head and neck cancer overall.

Table 3. Area Under the Receiver Operating Characteristic Curves (AUCs) of risk prediction models for head and neck cancer in hold-out testing set

Model	Men, AUC (95%CI)			Women, AUC (95%CI)		
	Head and neck cancer	Oral cavity cancer	Oropharyngeal cancer	Head and neck cancer	Oral cavity cancer	Oropharyngeal cancer
Epidemiological risk factors	0.69 (0.67-0.71)	0.69 (0.66-0.72)	0.66 (0.64-0.69)	0.75 (0.72-0.78)	0.75 (0.71-0.79)	0.76 (0.72-0.80)
Epidemiological risk factors and HPV serostatus			0.92 (0.90-0.94)			0.91 (0.86-0.94)
Epidemiological risk factors and PRS	0.72 (0.69-0.75)	0.73 (0.69-0.77)	0.71 (0.67-0.74)	0.75 (0.71-0.79)	0.79 (0.74-0.83)	0.76 (0.71-0.81)
Epidemiological risk factors, HPV serostatus and PRS			0.94 (0.92-0.95)			0.92 (0.88-0.95)

Epidemiological risk factor model includes age, smoking packyears, alcohol drinking intensity and education.
HPV, human papillomavirus; PRS, polygenic risk scores.

A risk prediction model for head and neck cancers incorporating lifestyle factors, HPV serology and genetic markers

Sanjeev Budhathoki, Brenda Diergaarde, Geoffrey Liu, Andrew Olshan, Andrew Ness, Tim Waterboer, Shama Virani, Patricia Basta, Noemi Bender, Nicole Brenner, Tom Dudding, Neil Hayes, Andrew Hope, Shao Hui Huang, Katrina Hueniken, Beatriz Kanterewicz, James D McKay, Miranda Pring, Steve Thomas, Kathy Wisniewski, Sera Thomas, Yonathan Brhane, Antonio Agudo, Laia Alemany, Areti Lagiou, Luigi Barzan, Cristina Canova, David I. Conway, Claire M. Healy, Ivana Holcatova, Pagona Lagiou, Gary J. Macfarlane, Tatiana V. Macfarlane, Jerry Polesel, Lorenzo Richiardi, Max Robinson, Ariana Znaor, Paul Brennan and Rayjean J. Hung

Supplementary Information

Supplementary Methods

Estimation of absolute risk

The absolute risk of developing head and neck cancer for an adult of age *a* years within a duration of τ years (i.e. within an interval of [*a, a+* τ]) was determined by integrating the equation below:

$$AR(a, a+\tau) = \int_{a}^{a+\tau} \lambda_0(t)\exp(Z\beta)\, exp\left(-\int_{a}^{t}[\lambda_0(u)\exp(Z\beta) + m(u)]du\right)dt$$

where $\lambda_0(t)$ is the baseline hazard function, *Z* is a set of risk factors, β is a vector of log relative risk, $m(t)$ is age-specific competing hazards of mortality, and u is the time interval for the estimation of the integral. The derivation of the equation has been described in detail elsewhere[1, 2]. The underlying assumption of the risk model is that risk factors act in a multiplicative fashion on the baseline hazard function. Odds ratios, estimated from cases and controls in our study with adjustment of age and other risk factors, were used as a measure of relative risk. The age-specific cancer rates and competing hazards for mortality (Supplemental Table 3) were obtained from Surveillance, Epidemiology, and End Results (SEER) Program and Centers for Disease Control and Prevention, National Center for Health Statistics database respectively[3, 4].

Model calibration

We evaluated calibration of the risk models in the UK Biobank cohort, which is a population-based prospective cohort study of over 500 000 participants. The details of the study design have been described previously[5]. In brief, participants of age ranging from 38 years to 73 years were recruited between 2006 and 2010 at multiple assessment centres across the United Kingdom. At baseline, all participants underwent a self-completed questionnaire survey which inquired about lifestyle risk factors such as smoking and alcohol use, and medical history and family history of cancer. In addition, extensive physical measurement and biospecimens were also collected at baseline. The information on cancer diagnosis was obtained through record linkage with death and cancer registries. For this study, participants were followed to the date of death, cancer diagnosis, or censoring date of March 31, 2016 (in England and Wales) and Oct 31, 2015 (in Scotland). A total of 481,881 participants were available for analysis including 749 cases of head and neck cancer. Genotyping was performed using the UK BiLEVE Axiom array and the UK Biobank Axiom array[6]. Imputation was based on the Haplotype Reference Consortium reference panel. We computed PRS in the UK Biobank using the same weights as in the model development set. Three variants [rs201982221, HLA-B (156-Trp), HLA-DRB1 (71-Glu)] were not genotyped or imputed in the UK Biobank and were not included in the calculation of PRS. We imputed serostatus of the UK Biobank participants by random binomial draw with the overall probability of seropositivity (0.86%) estimated from controls who were assayed in VOYAGER study.

UK Biobank is known to be a healthier population with higher social economic status, lower smoking rate and lower cancer incidence[7]. To account for the population-level difference in the risk profile in UK Biobank, we applied the recalibration approach with the models reported, using a random sample of 50% of the UK Biobank, while keeping the remaining 50% for strict prospective assessment of calibration. Recalibration is a standard statistical approach when a developed risk model is being imported into a population that may have different risk profiles, while keeping the model structure unchanged[8, 9]. The method details of recalibration have been reported previously[9, 10]. For our study, we computed the log-odds of HNC cancers (Z) in UKB based on the same coefficients of models we developed using the VOYAGER data. Then we fit a logistic regression

model in the 50% training sample with HNC cancer status as the outcome and Z as the sole predictor. The beta coefficient for Z, $\hat{\beta}_Z$, is the re-calibrated slope (i.e. the adjustment factor). The adjustment factors are summarized in Supplementary Table 7. The reported calibration is based on the 50% hold-out testing set. All absolute risk estimation and calibration analyses were performed in R statistical software using *iCARE* package.

Reference

1. Gail MH. Estimation and interpretation of models of absolute risk from epidemiologic data, including family-based studies. *Lifetime Data Anal* 2008;**14**: 18-36.

2. Pal Choudhury P, Maas P, Wilcox A, Wheeler W, Brook M, Check D, Garcia-Closas M, Chatterjee N. iCARE: An R package to build, validate and apply absolute risk models. *PLoS One* 2020;**15**: e0228198.

3. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Data, 9 Registries, Nov 2020 Sub (1975-2018) - Linked To County Attributes - Time Dependent (1990-2018) Income/Rurality, 1969-2019 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2021, based on the November 2020 submission.

4. Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death 1999-2019 on CDC WONDER Online Database, released in 2020. Data are from the Multiple Cause of Death Files, 1999-2019, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/ucd-icd10.html on Jul 19, 2021 12:15:53 PM.

5. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**: e1001779.

6. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**: 203-9.

7. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R, Allen NE. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017;**186**: 1026-34.

8. Field JK, Vulkan D, Davies MPA, Duffy SW, Gabe R. Liverpool Lung Project lung cancer risk stratification model: calibration and prospective validation. *Thorax* 2021;**76**: 161-8.

9. Puddu PE, Piras P, Kromhout D, Tolonen H, Kafatos A, Menotti A. Re-calibration of coronary risk prediction: an example of the Seven Countries Study. *Sci Rep* 2017;**7**: 17552.

10. Winter A, Aberle DR, Hsu W. External validation and recalibration of the Brock model to predict probability of cancer in pulmonary nodules using NLST data. *Thorax* 2019;**74**: 551-63.

Supplemental Table 1. Summary of the SNPs included in the calculation of polygenic risk score for head and neck cancer

	Variants	Region	Risk allele	Risk allele frequency	Odds ratio	Reference (PMID)
Oral cavity cancer						
	rs10462706	5p15.33	C	0.85	0.74	27749845
	rs1229984	4q23	G	0.94	0.57	27749845
	rs6547741	2p23.3	G	0.46	0.83	27749845
	rs8181047	9p21.3	A	0.24	1.24	27749845
	rs928674	9q34.12	G	0.12	1.33	27749845
Oropharyngeal cancer						
	rs1229984	4q23	G	0.94	0.55	27749845
	rs3828805	6p21.32	C	0.72	1.37	27749845
	rs4713462	6p21.3	A	0.326	0.71	34642315
	HLA-B*1501	6p21.3	P/A	0.059	0.79	34642315
	HLA-B (156-Trp)	6p21.3	P/A	0.061	0.80	34642315
	HLA-DRB1*1301	6p21.3	P/A	0.067	0.49	34642315
	HLA-DRB1 (71-Glu)	6p21.3	P/A	0.145	0.59	34642315
	HLA-DQA1*0103	6p21.3	P/A	0.078	0.53	34642315
	HLA-DQB1*0603	6p21.3	P/A	0.073	0.53	34642315
	rs35189640	12q23.3	T	0.02	1.66	34642315
Head and neck cancer†						
	rs1494961	4q21.23	C	0.49	1.12	21437268
	rs1789924	4q23	C	0.61	1.12	21437268
	rs4767364	12q24.13	A	0.30	1.13	21437268
	rs971074	4q23	G	0.88	0.75	21437268
	rs1229984	4q23	G	0.94	0.56	27749845
	rs1453414	11p15.4	C	0.2	1.19	27749845
	rs79767424	5p14.3	C	0.97	0.55	27749845
	rs2299187	7q21.11	A	0.02	3.26	27173062
	rs201982221	10q26	D/I	0.02	1.74	34642315
	rs35189640	12q23.3	T	0.02	1.79	34642315

D/I, deletion/insertion; P/A, presence/absence for amino acid polymorphisms in HLA alleles

†Including SNPs for oral cavity cancer and oropharyngeal cancer

Supplemental Table 2. Age-specific incidence rates of head and neck cancer and all-other-cause mortality rates per 100 000 person-years in non-Hispanic White population in the United States[a]

Age	Head and neck cancer		Oral cavity cancer		Oropharyngeal cancer	
	Incidence	All-other-cause Mortality	Incidence	All-other-cause Mortality	Incidence	All-other-cause Mortality
Men						
40-44	11.0	263.7	2.5	264.8	3.1	264.7
45-49	23.7	394.0	5.0	396.9	8.1	396.5
50-54	42.2	591.0	7.9	597.2	14.1	596.3
55-59	64.8	871.5	12.0	882.3	20.4	880.8
60-64	85.8	1278.4	15.5	1293.9	24.4	1292.2
65-69	101.5	1875.0	18.1	1894.6	24.8	1892.8
70-74	111.5	2906.1	20.0	2929.0	23.9	2927.4
Women						
40-44	4.2	154.6	1.1	155.0	0.8	155.0
45-49	8.2	235.0	2.3	235.8	1.9	235.8
50-54	13.6	352.4	3.7	354.0	3.4	353.9
55-59	20.6	524.8	5.7	527.5	5.2	527.3
60-64	26.7	792.8	7.5	796.6	6.6	796.4
65-69	33.6	1218.5	9.5	1223.7	7.5	1223.5
70-74	36.8	1968.4	11.4	1975.4	7.8	1975.4

[a]Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Data, 9 Registries, Nov 2020 Sub (1975-2018) - Linked To County Attributes - Time Dependent (1990-2018) Income/Rurality, 1969-2019 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2021, based on the November 2020 submission.

Supplemental Table 3. Distribution of the selected characteristics in cancer cases

Variables	Categories	Hypopharynx cancer	Larynx cancer	Other cancer*
Total (n)		518	2379	1077
Sex, n (%)				
	Men	438 (84.7)	2041 (85.8)	728 (67.7)
	Women	79 (15.3)	337 (14.2)	348 (32.3)
	Missing	1	1	1
Age (years), mean (SD)		61.4 (9.9)	63.4 (10.6)	58.7 (12.7)
Tobacco Smoking status, n (%)				
	Never	27 (6.2)	133 (6.4)	252 (26.4)
	Former	175 (39.9)	943 (45.6)	301 (31.5)
	Current	237 (54.0)	991 (47.9)	402 (42.1)
	Missing	79	312	122
Tobacco Pack-years, median (IQR)		40 (31.5)	42 (35.6)	36 (32.6)
Alcohol drinking status, n (%)				
	Never	53 (12.4)	396 (19.3)	201 (21.0)
	Former	217 (50.8)	958 (46.7)	371 (38.8)
	Current	157 (36.8)	699 (34.0)	385 (40.2)
	Missing	91	326	120
Drink/week, median (IQR)		28 (38.3)	21 (28.8)	14.7 (28.6)
Education, n (%)				
	Postsecondary	76 (20.9)	424 (23.6)	311 (34.0)
	High school diploma	88 (24.2)	436 (24.3)	275 (30.1)
	None/elementary	199 (54.8)	937 (52.1)	329 (36.0)
	Missing	155	582	162

*includes cancers of the salivary gland (C07.9-C08.9), nasopharynx (C11.0-C11.9) and oral cavity-oropharynx-hypopharynx not otherwise specified (C02.8, C02.9, C05.8, C05.9, C14.0, C14.2, C14.8).

Supplemental Table 4. Cut-off of smoking, drinking and polygenic risk score for head and neck cancers

	Variables	Categories	Head and neck cancer	Oral cavity cancer	Oropharyngeal cancer
Men					
	Smoking status[a]	Moderate	<24 pack-years		
		Heavy	≥24 pack-years		
	Drinking status[b]	Never/low	<5.5 drinks/week		
		Moderate	5.5 – <14.7 drinks/week		
		Heavy	≥14.7 drinks/week		
	Polygenic risk score[c]	1st tertile	≤ -0.08	≤ -0.27	≤ -0.43
		2nd tertile	> -0.08, ≤ 0.48	> -0.27, ≤ 0.0004	> -0.43, ≤ 0.21
		3rd tertile	> 0.48	> 0.0004	> 0.21
		Median (Q1, Q3)	0.23 (-0.28, 0.65)	-0.16 (-0.37, 0.03)	0.03 (-0.92, 0.31)
Women					
	Smoking status[a]	Moderate	<14 pack-years		
		Heavy	≥14 pack-years		
	Drinking status[b]	Never/low	<2.2 drinks/week		
		Moderate	2.2 – <6.9 drinks/week		
		Heavy	≥6.9 drinks/week		
	Polygenic risk score[c]	1st tertile	≤ -0.05	≤ -0.27	≤ -0.38
		2nd tertile	> -0.05, ≤ 0.54	> -0.27, ≤ 0.007	> -0.38, ≤ 0.27
		3rd tertile	> 0.54	> 0.007	> 0.27
		Median (Q1, Q3)	0.26 (-0.23, 0.68)	-0.16 (-0.37, 0.03)	0.06 (-0.78, 0.31)

[a]The cut-off is based on sex-specific medians among ever smokers in the control group

[b]The cut-off is based on sex-specific tertiles in the control group

[c]The polygenic risk scores are computed for oral cavity and oropharyngeal cancer separately based on the loci reported for these tumor types. Loci reported for head and neck cancer or their anatomical subsites are included in the PRS for head and neck cancer overall. The cut-off is based on sex-specific tertiles in the control group

Supplemental Table 5a. Beta coefficients of risk factors in different models of head and neck cancer overall and oral cavity cancer

	Men				Women			
	Epi model		Epi & PRS		Epi model		Epi & PRS	
	Estimate	P value	Estimate	P value	Estimate	P value	Estimate	P value
Head and neck cancer								
Age, < 50 years	0.44	<0.01	0.45	<0.01	0.59	0.03	0.62	0.02
50 - < 55 years	0.09	0.52	0.08	0.54	0.04	0.85	0.04	0.83
55 - < 60 years	0.22	0.11	0.19	0.16	-0.11	0.59	-0.11	0.57
60 - < 65 years	0.18	0.18	0.18	0.18	-0.08	0.70	-0.07	0.71
65 - < 70 years	-0.06	0.67	-0.07	0.63	-0.23	0.27	-0.24	0.24
70 - < 75 years	-0.06	0.72	-0.07	0.66	-0.19	0.38	-0.18	0.42
≥ 75 years	0.01	0.95	0.02	0.92	0.11	0.61	0.11	0.58
Smoking [a], Moderate	0.19	0.03	0.20	0.02	0.28	0.04	0.27	0.05
Heavy	0.91	<0.01	0.93	<0.01	1.24	<0.01	1.24	<0.01
Drinking [b], Moderate	-0.09	0.31	-0.14	0.10	-0.35	0.01	-0.36	0.01
Heavy	0.53	<0.01	0.49	<0.01	0.42	<0.01	0.41	<0.01
Education, High school	0.83	<0.01	0.81	<0.01	1.17	<0.01	1.15	<0.01
None/elementary	0.68	<0.01	0.67	<0.01	1.60	<0.01	1.58	<0.01
PRS category, 2nd tertile			0.42	<0.01			0.55	<0.01
3nd tertile			0.86	<0.01			0.66	<0.01
Oral cavity cancer								
Age, < 50 years	0.31	0.18	0.34	0.15	0.03	0.94	0.03	0.93
50 - < 55 years	-0.10	0.64	-0.10	0.63	-0.22	0.43	-0.23	0.41
55 - < 60 years	0.02	0.91	0.04	0.85	-0.39	0.15	-0.41	0.13
60 - < 65 years	0.08	0.70	0.10	0.63	-0.15	0.57	-0.19	0.47
65 - < 70 years	0.00	0.98	0.03	0.88	-0.07	0.80	-0.12	0.64
70 - < 75 years	0.29	0.18	0.33	0.14	-0.04	0.88	-0.02	0.94
≥ 75 years	0.64	<0.01	0.64	<0.01	0.43	0.09	0.44	0.09
Smoking [a], Moderate	0.42	0.01	0.43	0.01	0.35	0.05	0.31	0.08
Heavy	1.18	<0.01	1.18	<0.01	1.22	<0.01	1.21	<0.01
Drinking [b], Moderate	-0.08	0.55	-0.12	0.35	-0.31	0.10	-0.31	0.11
Heavy	0.65	<0.01	0.62	<0.01	0.45	<0.01	0.43	0.01
Education, High school	0.99	<0.01	1.00	<0.01	1.34	<0.01	1.33	<0.01
None/elementary	0.97	<0.01	0.99	<0.01	1.87	<0.01	1.86	<0.01
PRS category[b], 2nd tertile			0.29	0.02			0.55	<0.01
3nd tertile			0.77	<0.01			0.76	<0.01

[a]The cut-off is based on sex-specific medians among ever smokers in the control group

[b]The cut-off is based on sex-specific tertiles in the control group

Supplemental Table 5b. Beta coefficients of risk factors in different models of oropharyngeal cancer

	Men						Women					
	Epi model		Epi & HPV		Epi, HPV & PRS		Epi model		Epi & HPV		Epi, HPV & PRS	
	Estimate	P value	Estimate	P value	Estimate	P value	Estimate	P value	Estimate	P value	Estimate	P value
Age, < 50 years	0.74	<0.01	0.57	0.08	0.57	0.08	1.31	<0.01	1.57	0.01	1.60	0.01
50 - < 55 years	0.50	0.01	0.24	0.42	0.24	0.41	0.80	0.03	0.48	0.36	0.51	0.34
55 - < 60 years	0.67	<0.01	0.53	0.07	0.55	0.06	0.81	0.02	0.93	0.05	0.96	0.05
60 - < 65 years	0.53	<0.01	0.27	0.36	0.31	0.30	0.50	0.15	0.79	0.10	0.83	0.09
65 - < 70 years	0.34	0.08	0.49	0.10	0.49	0.10	0.18	0.63	0.40	0.44	0.44	0.40
70 - < 75 years	0.09	0.67	0.20	0.54	0.22	0.50	0.31	0.43	0.88	0.10	0.89	0.10
≥ 75 years	0.02	0.93	-0.03	0.95	-0.02	0.95	0.24	0.55	0.64	0.25	0.67	0.23
Smoking[a], Moderate	0.12	0.31	0.47	0.02	0.49	0.02	0.72	<0.01	1.02	0.01	1.04	<0.01
Heavy	0.75	<0.01	1.74	0.00	1.78	<0.01	1.26	<0.01	1.89	<0.01	1.89	<0.01
Drinking [b], Moderate	-0.10	0.38	-0.21	0.29	-0.24	0.22	-0.37	0.13	0.16	0.60	0.12	0.69
Heavy	0.49	<0.01	0.84	<0.01	0.80	<0.01	0.79	<0.01	1.06	0.00	1.05	<0.01
Education, High school	0.68	<0.01	0.57	<0.01	0.56	<0.01	0.69	<0.01	0.64	0.02	0.65	0.02
None/elementary	0.22	0.05	0.47	<0.01	0.50	<0.01	0.59	0.01	0.64	0.03	0.66	0.03
HPV seropositive			5.99	<0.01	5.96	<0.01			5.36	<0.01	5.32	<0.01
PRS category[b], 2nd tertile					0.11	0.51					-0.14	0.64
3rd tertile					0.50	<0.01					0.30	0.27

[a]The cut-off is based on sex-specific medians among ever smokers in the control group

[b]The cut-off is based on sex-specific tertiles in the control group

Supplemental Table 6. Odds ratios (ORs) and 95% confidence intervals (CIs) of head and neck cancer including and excluding HN5000 study

Variable	With HN5000		Without HN5000	
	OR (95% CI)	P value	OR (95% CI)	P value
Men				
Smoking status[a], Never	1 (Ref.)		1 (Ref.)	
Moderate	1.20 (1.01-1.43)	0.04	1.20 (1.00-1.45)	0.05
Heavy	2.58 (2.16-3.07)	<0.01	3.01 (2.51-3.62)	<0.01
Drinking status[b], Never/low	1 (Ref.)		1 (Ref.)	
Moderate	0.85 (0.72-1.01)	0.06	0.91 (0.76-1.09)	0.31
Heavy	1.57 (1.34-1.84)	<0.01	1.54 (1.30-1.84)	<0.01
Education, Postsecondary	1 (Ref.)		1 (Ref.)	
High school diploma	2.14 (1.83-2.52)	<0.01	2.55 (2.15-3.03)	<0.01
None/elementary	1.48 (1.20-1.82)	<0.01	3.24 (2.53-4.14)	<0.01
Polygenic risk score[b], 1st tertile	1 (Ref.)		1 (Ref.)	
2nd tertile	1.52 (1.29-1.79)	<0.01	1.40 (1.17-1.68)	<0.01
3rd tertile	2.34 (2.00-2.75)	<0.01	2.18 (1.83-2.58)	<0.01
Women				
Smoking status[a], Never	1 (Ref.)		1 (Ref.)	
Moderate	1.31 (1.00-1.70)	0.05	1.21 (0.91-1.60)	0.19
Heavy	3.67 (2.91-4.64)	<0.01	3.65 (2.86-4.65)	<0.01
Drinking status[b], Never/low	1 (Ref.)		1 (Ref.)	
Moderate	0.62 (0.47-0.82)	<0.01	0.81 (0.61-1.08)	0.15
Heavy	1.34 (1.05-1.70)	0.02	1.12 (0.86-1.46)	0.40
Education, Postsecondary	1 (Ref.)		1 (Ref.)	
High school diploma	2.75 (2.17-3.48)	<0.01	3.18 (2.49-4.05)	<0.01
None/elementary	2.31 (1.68-3.17)	<0.01	4.91 (3.34-7.22)	<0.01
Polygenic risk score[b], 1st tertile	1 (Ref.)		1 (Ref.)	
2nd tertile	1.60 (1.24-2.05)	<0.01	1.67 (1.28-2.18)	<0.01
3rd tertile	1.85 (1.45-2.37)	<0.01	1.78 (1.37-2.31)	<0.01

OR, odds ratio; CI, confidence interval

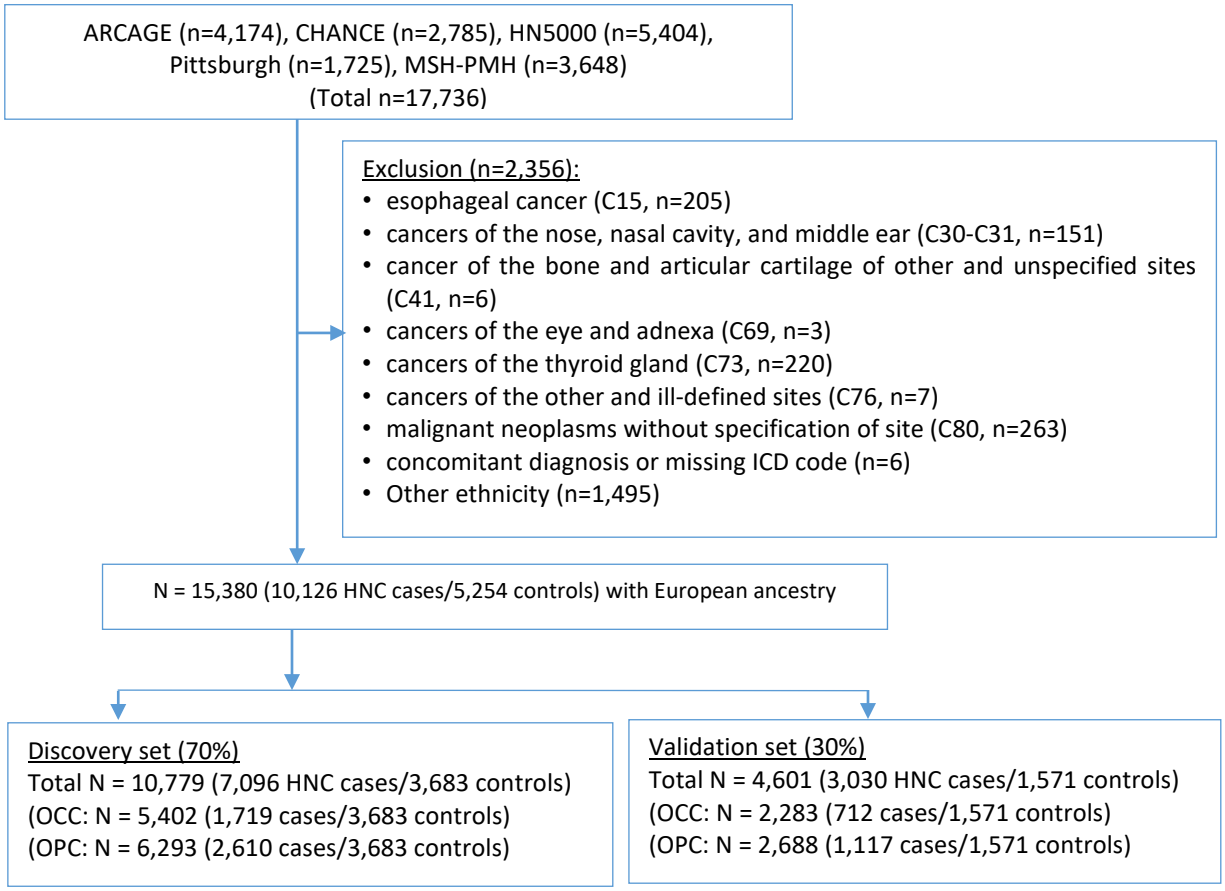
[a]The cut-off is based on sex-specific medians among ever smokers in the control group

[b]The cut-off is based on sex-specific tertiles in the control group

Supplemental Table 7. Adjustment factors ($\hat{\beta}_Z$) for UKB

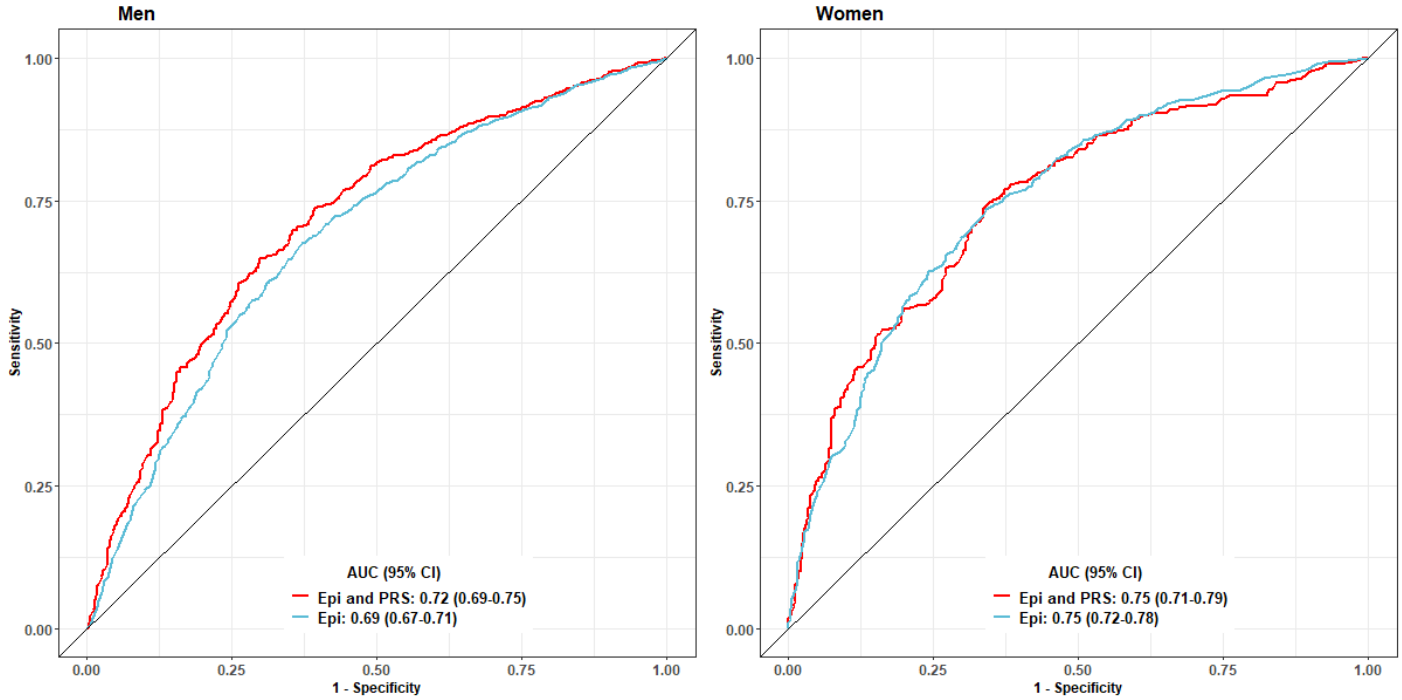
	Men	Women
Head and neck cancer	0.737	0.407
Oral cavity cancer	0.630	0.306
Oropharyngeal cancer	0.830	0.890

Supplemental Figure 1. Flowchart of the study subjects

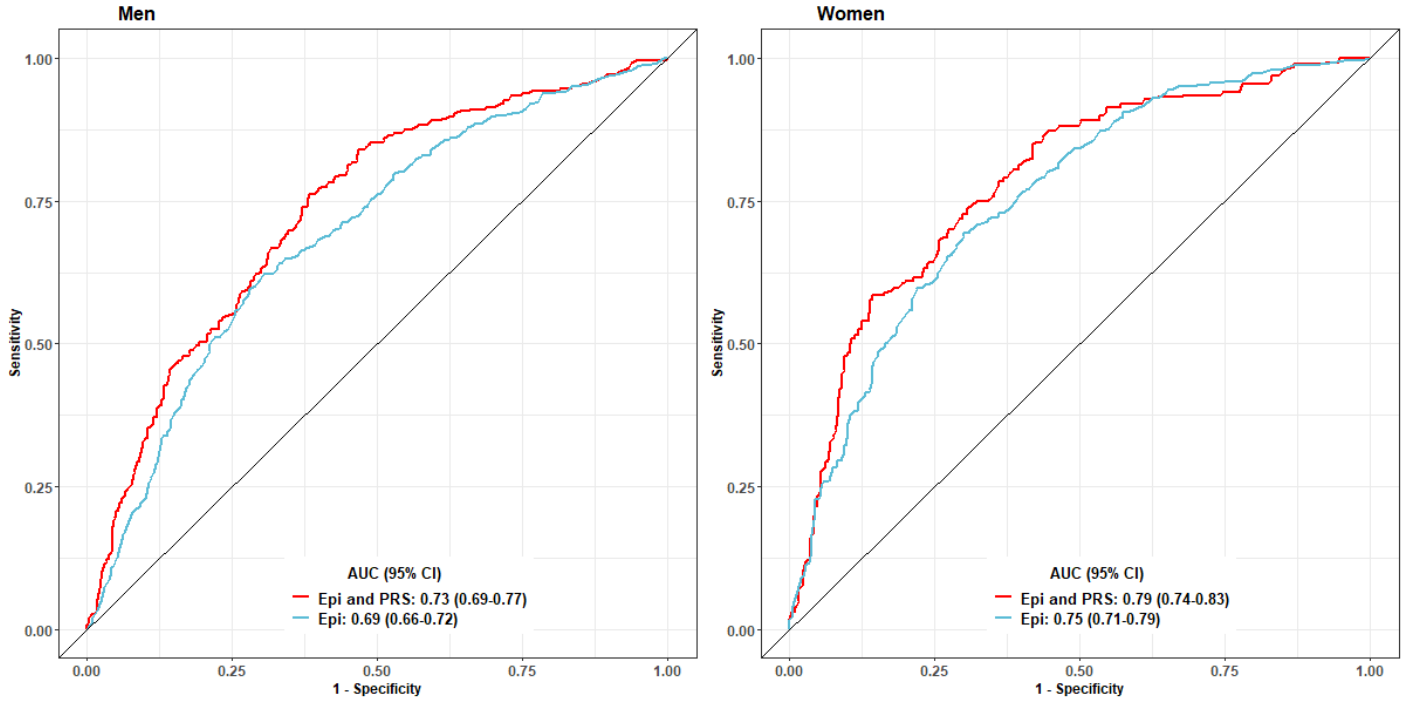


Supplemental Figure 2. Receiver Operating Characteristic Curves (ROCs) of risk models for head and neck cancer in hold-out testing set

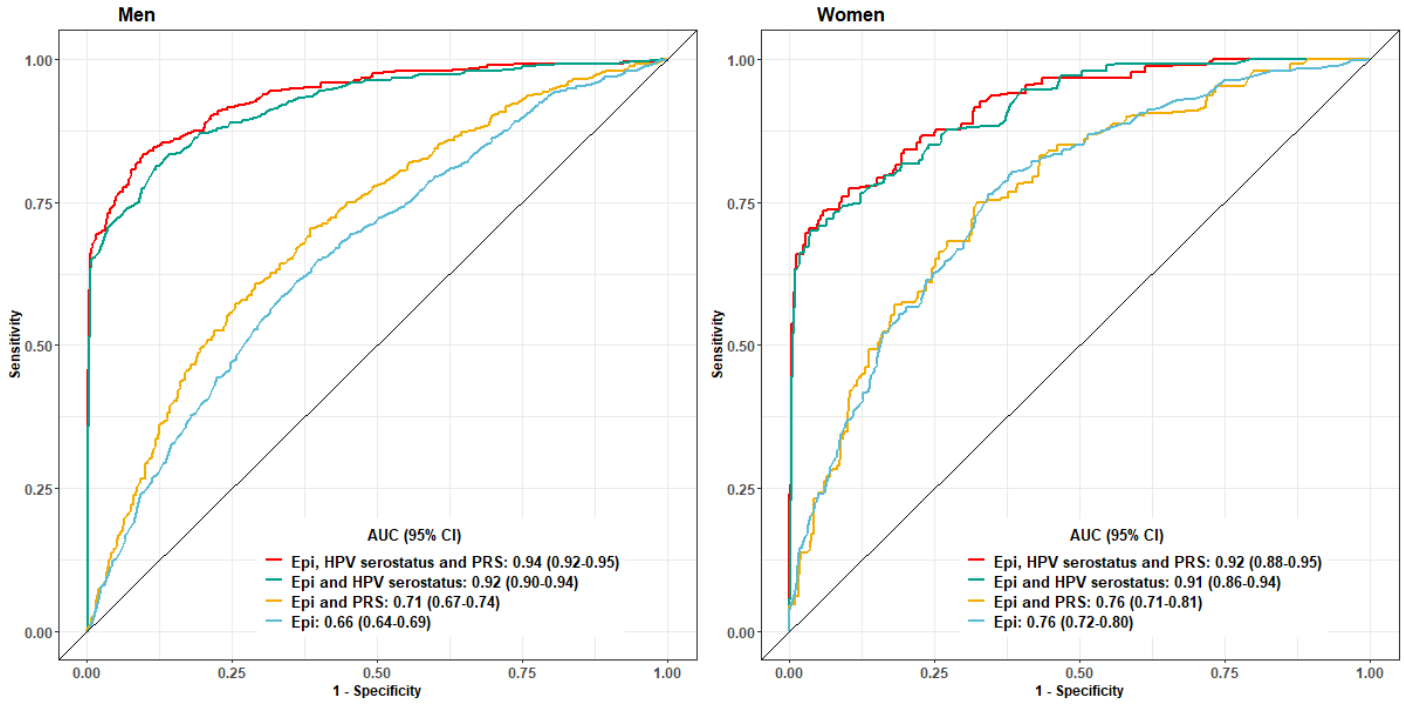
A. Head and neck cancer



B. Oral cavity cancer



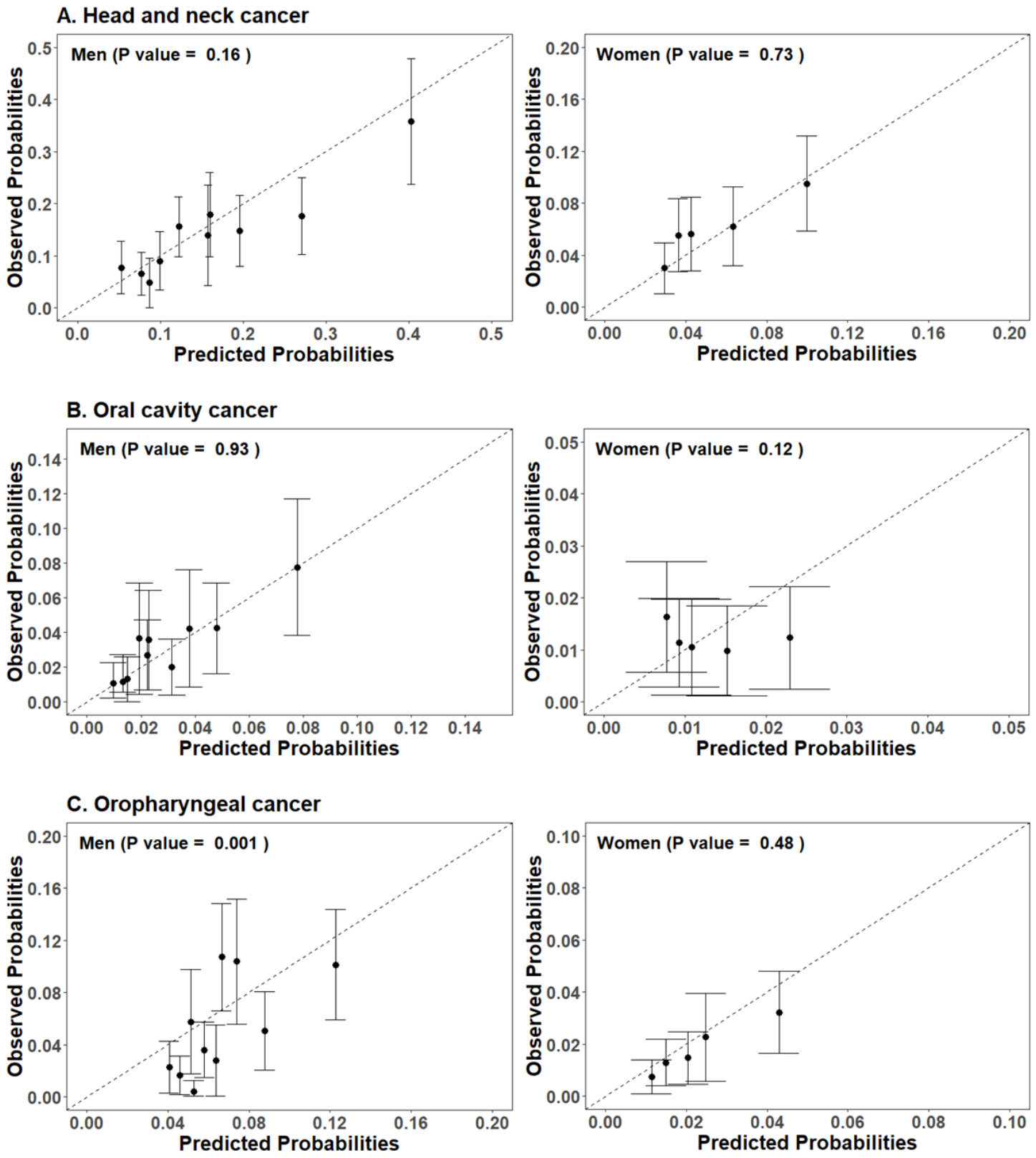
C. Oropharyngeal cancer



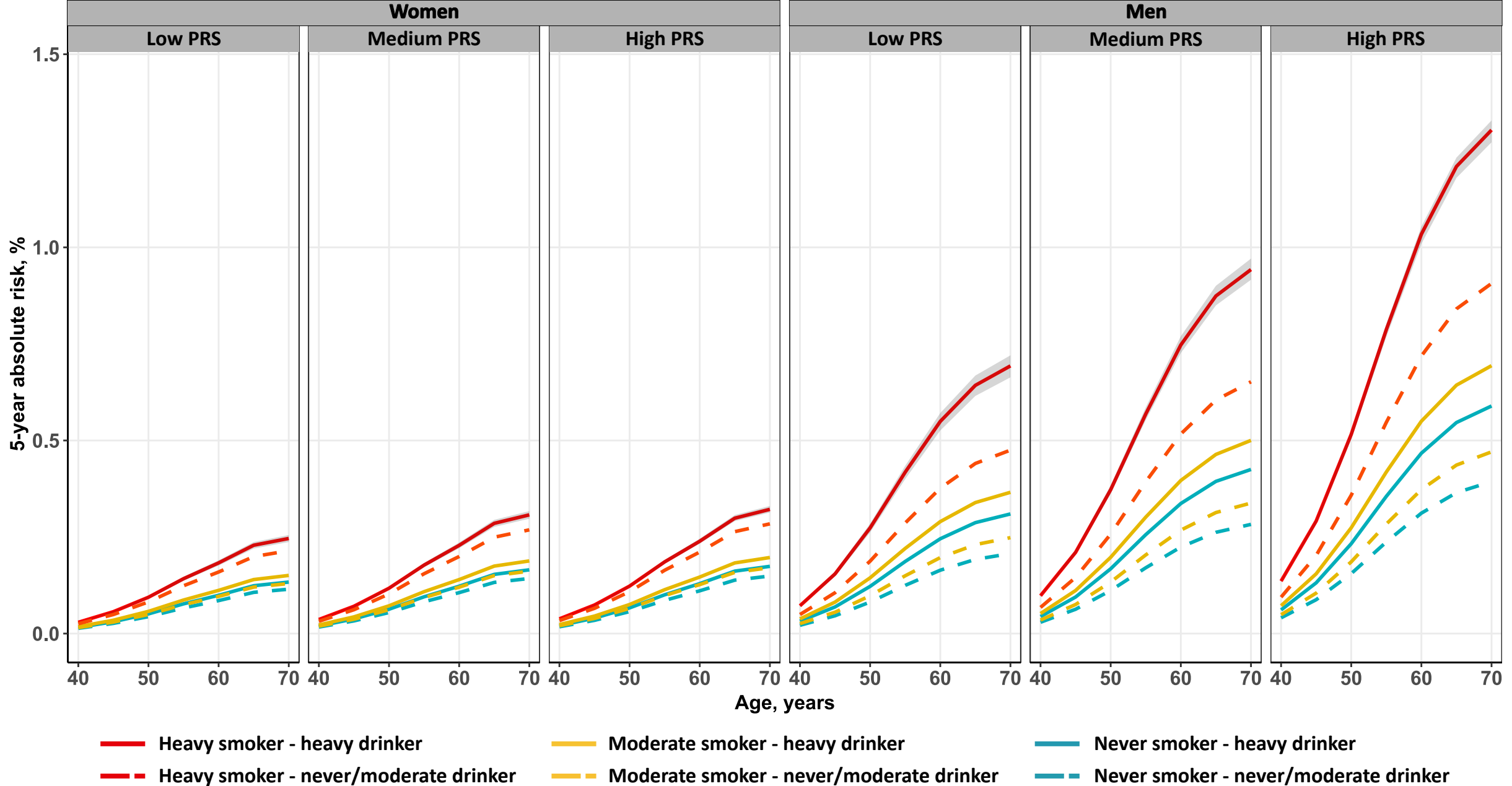
HPV, human papillomavirus; PRS, polygenic risk scores.

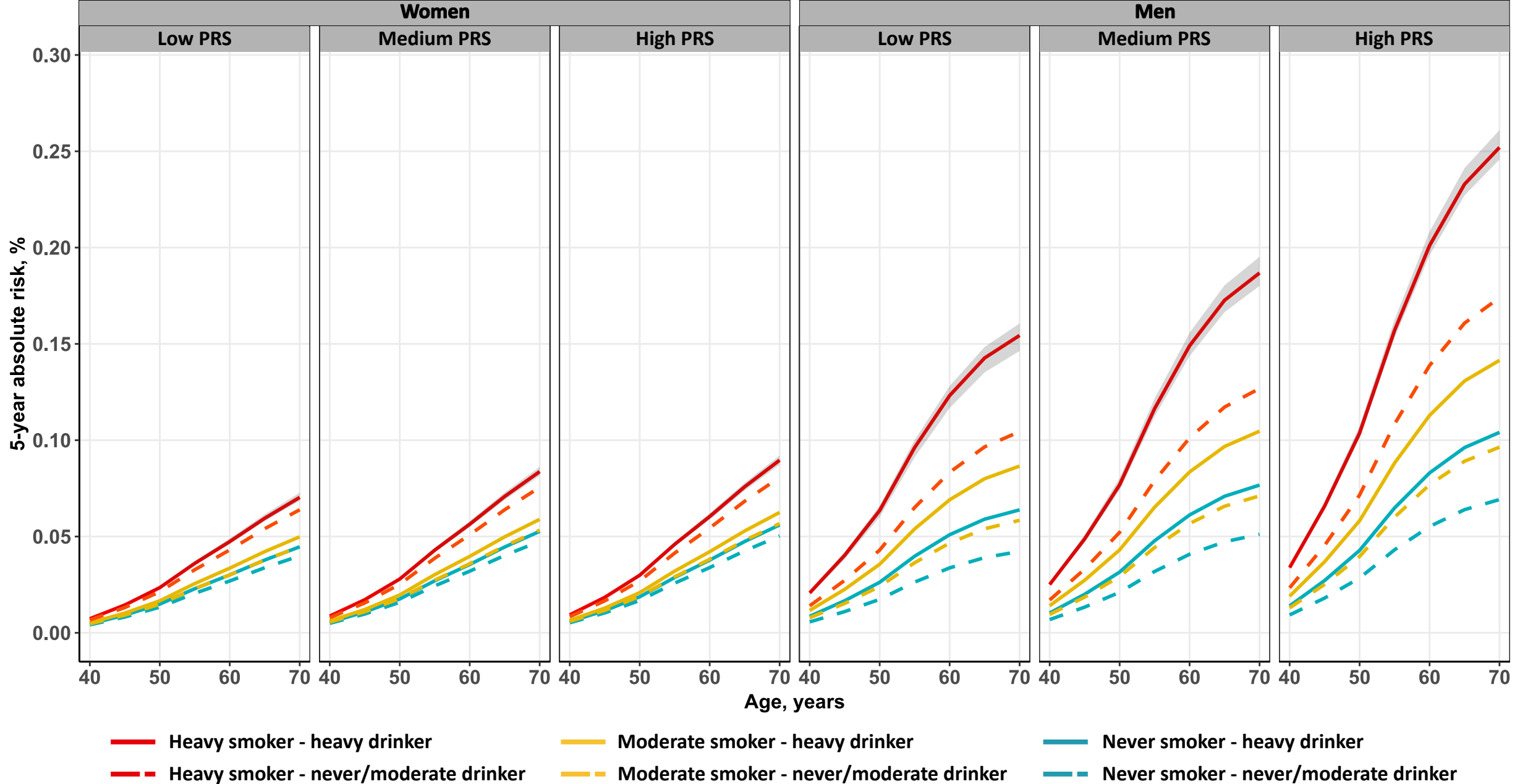
Epidemiological (epi) risk factor model includes age, smoking packyears, alcohol drinking intensity and education. The model for head and neck cancer overall (**A**) and oral cavity cancer (**B**) include epidemiological risk factors and polygenic risk score. The model of oropharyngeal cancer (**C**) includes epidemiological risk factor, HPV serostatus and polygenic risk score. The left and right panel shows the ROC curves of risk models for head and neck cancer in men and women, respectively.

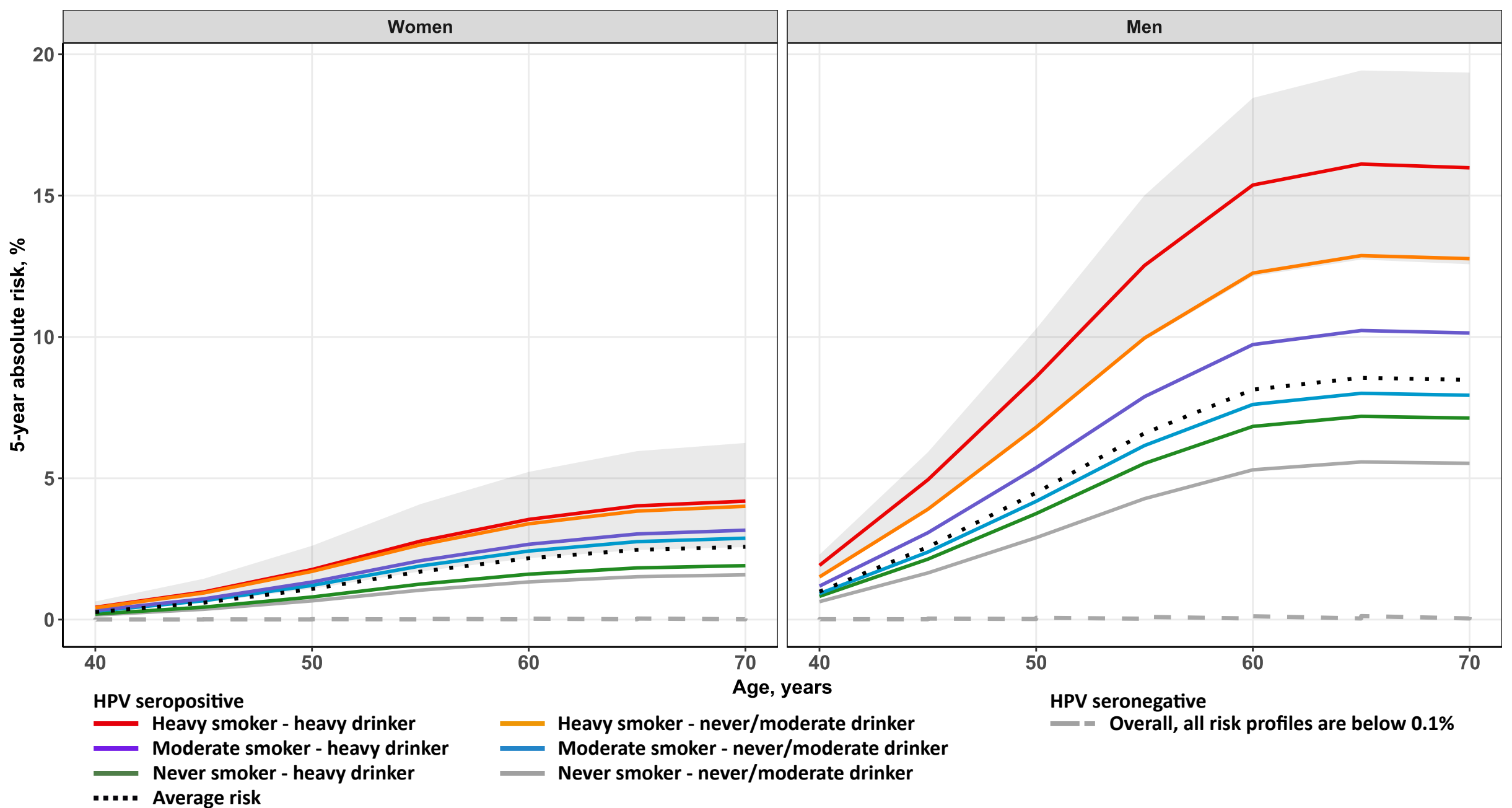
Supplemental Figure 3. Calibration plot comparing predicted probability with observed probability.



The model for head and neck cancer overall (**A**) and oral cavity cancer (**B**) include epidemiological risk factors and polygenic risk score. The model of oropharyngeal cancer (**C**) includes epidemiological risk factor, HPV serostatus and polygenic risk score. The calibration lines for men (Left panel) are plotted in deciles of predicted probability and for women (Right panel) are plotted in quintile due to smaller sample size. P-values are based on Hosmer-Lemeshow test.


Women
Men
Low PRS
Medium PRS
High PRS
Low PRS
Medium PRS
High PRS
5-year absolute risk, %
1.5
1.0
0.5
0.0
40
50
60
70
Age, years
Heavy smoker - heavy drinker
Heavy smoker - never/moderate drinker
Moderate smoker - heavy drinker
Moderate smoker - never/moderate drinker
Never smoker - heavy drinker
Never smoker - never/moderate drinker


Women
Men
Low PRS
Medium PRS
High PRS
Low PRS
Medium PRS
High PRS
5-year absolute risk, %
0.30
0.25
0.20
0.15
0.10
0.05
0.00
40
50
60
70
Age, years
Heavy smoker - heavy drinker
Heavy smoker - never/moderate drinker
Moderate smoker - heavy drinker
Moderate smoker - never/moderate drinker
Never smoker - heavy drinker
Never smoker - never/moderate drinker


Women
Men
5-year absolute risk, %
20
15
10
5
0
40
50
60
70
Age, years
HPV seropositive
Heavy smoker - heavy drinker
Heavy smoker - never/moderate drinker
Moderate smoker - heavy drinker
Moderate smoker - never/moderate drinker
Never smoker - heavy drinker
Never smoker - never/moderate drinker
Average risk
HPV seronegative
Overall, all risk profiles are below 0.1%