Combining Machine Learning with Bayesian Inverse Modelling to Estimate the Conditional Probability of Developing Oropharyngeal Cancer following an Oral Human Papilloma Virus Infection

*Prerna Tewari1, 2, *Eugene Kashdan3, Cathal Walsh4, Cara M Martin1, 2 Andrew C Parnell5, John J O’Leary1, 2

1Dept. Histopathology and Morbid Anatomy, Trinity College Dublin, Dublin, Ireland
2Dept. Pathology, Coombe Women & Infants University Hospital, Dublin, Ireland
3College of Business, University College Dublin, Dublin, Ireland
4Dept. Mathematics and Statistics, University of Limerick, Limerick, Ireland
5Hamilton Institute, Insight Centre for Data Analytics, Maynooth University, Kildare, Ireland

*Equal contribution

Corresponding authors:
Prerna Tewari
Dept. Histopathology and Morbid Anatomy,
Trinity College Dublin/Coombe Women and Infants University Hospital,
Dublin 8, Ireland
Email: tewarip@tcd.ie

Eugene Kashdan
College of Business
University College Dublin
Dublin 4, Ireland
Email: eugene.kashdan@ucd.ie

Keywords: Human Papillomavirus, Oropharyngeal cancer, Machine learning
Abstract

Background

Despite an epidemic increase in the prevalence of Human Papilloma Virus (HPV) related Oropharyngeal Squamous Cell Carcinomas (OPSCCs) in Northern America and parts of Europe, there is virtually no information about the natural history of these cancers. The lack of well-defined precursor lesions and limited data on oral HPV persistence and clearance rates, poses a challenge for disease modelling. We propose a novel mathematical modelling approach to estimate the conditional probability of developing HPV related OPSCCs following a prevalent HPV infection and other covariates.

Methods

We developed a double-Bayes method, whereby a Bayesian machine learning model first estimates the probability of an individual having an oral HPV infection, given OPSCC and other covariate information. The model is then inverted using Bayes’ theorem to reverse the probability relationship. The mathematical model was derived from two datasets representing the adult population in the United States (US), the Surveillance Epidemiology and End Results Program (SEER) Head and Neck with HPV Status Database and the National Health and Nutrition Examination Survey (NHANES) 2011-2014.

Results

The model dataset contains 8,623 subjects of which 70.7% had a prevalent oral HPV positive infection. When stratified by age, sex, marital status and race/ethnicity, the model estimated higher conditional probability for developing OPSCCs following an oral HPV infection in non-Hispanic White males and females compared to other race/ethnicities. Non-Hispanic White males with an oral HPV infection had nearly two fold higher risk of developing OPSCC than non-Hispanic White females (10.6 cases per 10,000 thousand vs 5.05 cases per 10,000) in the age range 50-60.
Conclusion

We have employed a novel statistical approach to estimate the conditional probability of developing OPSCCs following an oral HPV infection and covariates age, sex, ethnicity and marital status in the US population. We recognise that at best this is a first guess estimate of a natural history model of HPV driven OPSCCs within the existing limitations of the model.
Background

HPV infections are now firmly established as the primary cause of cervical cancers, some anogenital cancers, and a subset of Head and Neck cancers [1-3]. Head and neck cancers represent a diverse group of cancers that develop from different anatomical sites; oral cavity, oropharynx (tonsils, base of tongue), nasal cavity, nasopharynx, hypopharynx and larynx [4]. While there has been a decrease in the incidence of laryngeal, hypopharyngeal and oral cavity cancers over the last few decades, due to decrease in smoking rates, a sharp increase in the incidence of oropharyngeal squamous cell carcinomas (OPSCCs) has been observed over the same period in several developed countries worldwide [5]. The increase in OPSCC incidence rates has been causally linked to the growing prevalence of HPV infections presumably acquired via increased oral exposure to infected anogenital sites with changing sexual behaviour [6].

In the United States (US), the annual incidence rate of OPSCCs has now surpassed that of cervical cancer, with similar trends projected for other developed countries, there is a growing concern for managing disease burden [7]. While cervical cancer has been managed effectively by well-established public health measures including screening and prophylactic HPV vaccination, similar efforts have yet to be established for OPSCCs [8, 9].

Unlike cervical cancer, precursor lesions are not associated with OPSCCs, but it is likely that a subclinical HPV infection that persists for decades precedes the development of these cancers similar to cervical cancer. The key parameters that govern the natural history of oral HPV infection (transition from infection to malignancy) remain largely ill-defined because they cannot be easily inferred from experimental data. While several population based studies have reported oral HPV infections prevalence rates at 4-7.0% in healthy populations with similar associated risk factors as the ones reported for OPSCCs including gender, sexual behaviour and current tobacco use, limited data exists on oral HPV persistence and clearance rates and associated risk factors [10, 11].

Unravelling the trajectory of oral HPV infections and associated risk factors that promote malignant transformation is crucial to the development of successful primary and secondary prevention strategies.
Mathematical models have been previously been used successfully to estimate some of these ill-defined parameters in cervical cancer [12]. A comprehensive mathematical model incorporating data on oral HPV prevalence, persistence and/or clearance rates as well as demographics, lifestyle risk factors and sexual history will help to delineate the natural history of HPV in OPSCCs and associated risk factors.

Our aim is to estimate the conditional probability of developing OPSCCs following an oral HPV infection along with socio-demographic and other covariate information. To achieve this, we developed a novel modelling approach, which calibrates a machine learning method using inverse conditional probabilities. We made use of the Bayesian Additive Regression Trees (BART) machine learning approach as it allows for complex non-linear interactions between variables and produces probabilistic confidence and prediction intervals. We stress that our approach is a proof-of-concept and outline some possible extensions.

Materials and Methods

Datasets

The first dataset is from the Surveillance, Epidemiology, and End Results (SEER) program which provides information on cancer statistics in the U.S. population [13]. Since 2010, SEER has collected data on the oral HPV status of Head and Neck Cancer patients https://seer.cancer.gov/seerstat/databases/hpv. Data is available for patients diagnosed between 2010 and 2016 in 12 SEER registries. The oral HPV status information has been recoded as: 1) HPV Negative; 2) HPV Positive; 3) Unknown/NA. The data also includes individual-level and aggregated county-level demographic and socio-economic information.

The second dataset we use for the analysis was provided by the National Health and Nutrition Examination Survey (NHANES) http://www.cdc.gov/nchs/nhanes.htm. NHANES is a stratified, multistage, clustered probability sample that is representative of the noninstitutionalized, civilian US population [14]. NHANES includes a series of surveys as well as laboratory tests and detailed questionnaires.
Since 2011, NHANES has been collecting data on oral HPV infections in the US population. Participants provided a 30-second oral rinse and gargle with mouthwash. DNA from oral exfoliated cells was evaluated by polymerase chain reaction and type-specific hybridization for HPV detection.

**Data Wrangling and Exploratory Data Analysis**

The oral HPV data and demographic characteristics of the Head and Neck cancer patients was extracted from the specialised “Head and Neck with HPV Status” SEER database (access granted on request). Only those cases that were classified as oropharynx based on ICD-O-3 site codes with a confirmed oral HPV status were considered for data modelling.

The NHANES data is stored in SAS format as a separate file for each group of variables and each period of the survey. The datasets for 2011-2012 and 2013-2014 surveys provided the most granular demographic data and were combined for analysis. The files were imported into R and linked through the unique respondent ID. The data for minors (14-17 years old) is restricted and not considered for the analysis.

Since, we propose to combine the SEER and NHANES datasets together in a Bayesian probability calculation, we require them to have variables which can be matched together. Unfortunately, this necessitates removing all variables which cannot be similarly coded. The variables that remain after recoding are: age, sex, marital status, and race/ethnicity (non-Hispanic Blacks, non-Hispanic Whites, Hispanics and other races including American Indian/Alaska Native, Asians and Pacific Islanders).

**Data Modelling**

To estimate the conditional probability of developing OPSCCs following an oral HPV infection, we use a novel double-Bayesian method whereby a Bayesian machine learning model is first used to estimate the probability of an individual having an oral HPV infection, given OPSCC and other covariate information from the SEER data. This model is then inverted using Bayes’ theorem to reverse the probability relationship. The inversion involves corrections using incidence values which are obtained from both SEER and NHANES data. We have made our code available at www.github.com/andrewcparnell/HNSCC for
those wishing to confirm or extend our analysis. However we note that the SEER oral HPV data are restricted and not included in the Github repository.

**Notation**

We define $y$ as the event that an individual has OPSCC and $z$ as the event that an individual has an oral HPV infection. We define $x_j$ to be the covariate values for an individual on covariate $j$, with $j = 1, ..., M$ covariates, and write $x$ to be the set of all covariates for an individual. In practice $M = 4$ with $x_j$ representing, respectively, age, sex, marital status and race/ethnicity.

**Overarching Bayesian Framework**

We use Bayes’ theorem to invert the probabilities and obtain our desired goal $P(y|z,x)$ for an individual. That is, the probability that an individual has OPSCC given that they have an oral HPV infection and covariate values $x$. To estimate $P(y|z,x)$ directly we would require a longitudinal data set for which all individuals had an oral HPV infection and only a subset developed OPSCC.

We thus calculate:

$$P(y,z|x) = \frac{P(z|y,x)P(y|x)}{P(z|x)} \quad (i)$$

where now $P(z|y,x)$ is the probability of an individual getting HPV given they have OPSCC, $Pr(y|x)$ is the probability of getting OPSCC given covariates $x$, and $Pr(z|x)$ is the probability of getting HPV given covariates $x$. We discuss how we obtain numerical estimates for each of these in turn below but, briefly, $P(z|y,x)$ is obtained from the SEER data via our BART model, $Pr(y|x)$ is obtained from SEER incidence estimates, and $Pr(z|x)$ is obtained from NHANES incidence estimates of oral HPV.

**BART**

The predictive probabilities $P(z|y,x)$ were calculated from the SEER data by employing BART. We use the relationship:
\[
Pr(z|y, x) = \int p(z, \theta|y, x, D) \, d\theta \quad \text{(ii)}
\]

where \(\theta\) are a set of parameters (detailed below) arising from posterior distribution calculated by the BART model, and \(D\) is the SEER data set consisting of triples \((x_i, y_i, z_i)\) for individuals \(i = 1, \ldots, 8623\).

The probability density \(p(z, \theta|y, x, D)\) is created from:

\[
p(z, \theta|y, x, D) = p(\theta|D)p(z|\theta, y, x, D) \quad \text{(iii)}
\]

where \(p(\theta|D)\) is the BART posterior distribution and \(p(z|\theta, y, x, D)\) is the predictive distribution of oral HPV for a new individual with OPSCC status \(y\) and covariate value \(x\).

The BART posterior distribution is built on a relationship that estimates the probability of oral HPV occurrence from OPSCC and covariates via a latent probit model [15], where first:

\[
p(\theta|D) \propto p(\theta) \prod_{i=1}^{8623} p(y_i|x_i, z_i, \theta) \quad \text{(iv)}
\]

where \(p(\theta)\) is a prior distribution on a set of parameters \(\theta\) and \(\prod p(y_i|x_i, z_i, \theta)\) is a likelihood term. In a BART probit classification model the likelihood is structured so that, at the top level:

\[
y_i = \begin{cases} 
1 & \text{if } \tilde{y}_i > 0 \\
0 & \text{otherwise}
\end{cases}
\]

where \(\tilde{y}\) is a latent parameter which is given a normally distributed likelihood:

\[
\tilde{y}_i \sim N(\mu_i, 1)
\]

The key component of BART is that the \(\mu_i\) parameters are set to be a sum of regression trees, i.e.:

\[
\mu_i = \sum_{j=1}^{J} g_j(x_i, z_i, \theta_j) \quad \text{(v)}
\]

with \(\theta_j\) controlling the leaf node parameters of tree \(j\). Each tree is made up of a set of decision nodes (based on \(x_i\) and \(z_i\)) to provide a prediction \(\theta_j\) which, when summed across trees, produces an estimate of \(\tilde{y}_i\) which is then compared with \(y_i\) through the equation above. The full set of parameters \(\theta\) contains all the tree
decision parameters $g_j$, the prediction parameters $\theta_j$ and the latent normal parameters $\tilde{\mu}_i$ for all observations.

The model fitting stage of BART involves first guessing at values of the parameters and updating them using Markov Chain Monte Carlo [16] to produce a large set of estimates of the parameters. It can be shown that the latent parameter formulation above can be re-expressed as a binomial probit model with:

$$y_i | x_i, z_i \sim Bin(1, \phi(\mu_i)) \quad (vi)$$

where $\phi$ is the standard normal cdf. We use the posterior estimates of $\phi(\mu)$ as probabilities of oral HPV given OPSCC.

We run the probit BART model using the bartMachine package, which by default uses $J = 50$ trees and produces 1000 posterior samples of the parameters [17].

**Bayesian Inversion with BART**

Our approach, taken together, involves the following steps:

1. Fit the BART probit model in equation (vi) to the SEER data set to estimate the probability of an oral HPV infection given OPSCC and covariates for any given adult.

2. For each desired new set of values $x$ comprising a set of covariate values:
   a. Simulate from the posterior distribution a predicted probability of oral HPV given OPSCC for these covariate values.
   b. Look up the SEER and NHANES incidence rates for that combination of covariates $x$.

These are assumed known without uncertainty.
c. Compute the probability of OPSCC given oral HPV and covariates using equation (i). This involves multiplying the simulated probability from step (a) above by the ratio of the incidence rates.

3. Repeat step 2 to form a posterior distribution of probabilities.

We subsequently summarise these probabilities to produce 95% posterior uncertainty intervals.

Results

Study Population
The model dataset contains 8,623 individuals with a mean age of 60.9 years. Males accounted for 82.9% of the population with an oral HPV infection rate of 73.6%. Females constituted 17.1% of the entire population and had an oral HPV infection rate of 56.4%. Amongst the different race/ethnic groups, non-Hispanic Whites represented 80.1% of the entire dataset with an oral HPV prevalence rate of 74.4%. Characteristics of the study participants included in the model are presented in Table 1.

The NHNAES study population included 9,134 respondents with a confirmed oral HPV data. The mean age of participants was 42.1 years. Males comprised 49.2% of the population with an oral HPV prevalence rate of 11.9% and females represented 50.8% with an oral HPV prevalence rate of 3.8%. Demographic details and oral HPV prevalence rates are summarised in Table 2.

Performance of the BART Model
The BART model was run on the SEER HPV data using the default 50 trees and 1000 posterior iterations (removing a proportion for a warm-up period). We use 75% of the data for training purposes and the remaining 25% constituted the test data set with similar proportions of HPV negative/positive as the full dataset.

We evaluated the performance of the model via the Receiver Operator Characteristic (ROC) curve and the Area Under the Curve (AUC) value. Using only age, sex, marital status and race/ethnicity as covariates, we
obtained an AUC value of 0.70. We also had the possibility of using further covariates such as smoking and employment status but these are only available at an aggregated level; no model we tried increased the AUC beyond 0.70. In the end we removed these extra covariates for simplicity in the final interpretation of the model.

**Conditional Probability of Developing OPSCC given Oral HPV by Age, Sex, Marital Status and Race/Ethnicity**

The estimated probabilities of an individual developing OPSCC following an oral HPV infection are shown in Figure 1. We note that due to the nature of the machine learning model, all probabilities are calculated on the full set of covariates: age, sex, marital status and race/ethnicity.

The risk estimates for males show an increased probability for non-Hispanic White males up to the age of 60 (10.6 cases per 10,000), with lower risk estimates for Hispanics (5.6 cases per 10,000) and non-Hispanic Blacks (4.6 per 10,000). The risk estimates for females show similar trends with an increased probability for non-Hispanic White females (4.95 cases per 10,000) of developing OPSCCs, whilst risk estimates for Non-Hispanic Blacks and Hispanics are broadly similar for the age range 40-50. It appears that Hispanic females have a lower chance of developing OPSCC than other ethnicities, but there is likely considerable uncertainty in these values due to lower sample sizes.

**Conditional Probability of Developing OPSCC given HPV by Age, Sex and Marital Status for Non-Hispanic Whites**

The estimated probabilities of an individual developing OPSCC given they have HPV for married Non-Hispanic Whites by age is shown in Figure 2. Whilst there is a clear increase across ages for both sexes, there appears to be little difference in the probabilities of developing OPSCCs for non-Hispanic White males and females up to the age of 50. For the age range 50-60, males have a substantially higher risk of developing OPSCCs than females (10.6 cases per 10,000 vs 5.05 cases per 10,000) following an HPV
infection. However, we observe a substantial increase in the number of cases for women >60 and a decline in the number of OPSCC cases in males aged >60 and above.

Discussion

We have developed a novel mathematical modelling approach – a “double Bayes” that inverts the results of a machine learning BART method using Bayes Theorem to estimate the correct conditional probabilities for developing OPSCCs following a prevalent oral HPV infection and other associated covariates: age, sex, marital status and race/ethnicity in the US population. This approach represents the first step in delineating the natural history of HPV related OPSCCs, a disease notoriously difficult to model given the lack of well-defined histological end-points and associated HPV infections.

The model estimates a substantially higher risk of HPV related OPSCCs for married non-Hispanic White males and females when compared to other races/ethnic groups. We appreciate that there is but likely considerable uncertainty in the probability estimates due to lower sample sizes, nonetheless similar risk estimates have been observed in previous cohort studies on HPV related OPSCCs, although with the caveat that the risk estimates we report are derived from Bayesian modelling [5, 18].

Despite the mathematical novelty of our approach, and in particular the ability to invert the association between OPSCC and HPV, we foresee several difficulties and issues with the results we present. The first is that we have no means to estimate the time interval between an individual having an oral HPV infection and subsequent development of OPSCC as no longitudinal follow-up data was available in either of the datasets. The NHANES dataset is a cross-sectional dataset with a single time point measure of oral HPV infections in healthy individuals. The lack of longitudinal follow-up data prevents estimation of oral HPV persistence, clearance rates and the potential role in malignant transformation. While the SEER dataset provides no data on oral HPV infection in individuals before the diagnosis of OPSCC. An ideal data set would collect both time of diagnosis of oral HPV and OPSCC, and use a double censoring model to estimate the time lag similar to HIV natural history models [19]. A second issue relates to the uncertainty estimates
for the incidence of OPSCC or HPV. If these were available, we could propagate them through our inversion technique to produce more conservative uncertainty intervals. Currently we believe our results, which we stand over as a best first guess, underestimate the interval widths at any given confidence level.

A third limitation is the cross-categorisation of various covariates collected as part of the analysis. Between SEER and NHANES, there are different levels of categorisation of both age and race/ethnicity which requires some level of re-coding and so weakens the final probabilistic predictions we create. Furthermore, due to limited /lack of data on smoking, sexual behaviour and sexual history on individuals in the SEER dataset, we were unable to include these known risk factors for HPV related OPSCCs as covariates in our model [20].

The major strength of the proposed modelling approach lies in developing and validating the oral HPV natural history model in a large sample size of participants, pooled through two well characterised datasets; NHANES and SEER, representing healthy individuals and OPSCC patients derived from the same population.

**Conclusion**

In summary, we have developed and described a novel statistical modelling approach to estimate the conditional probability of developing OPSCCs following an HPV infection and covariates age, sex, marital status and race/ethnicity in the US population. We appreciate the limitations of the proposed model and at best recognise that this is a first guess estimate of a natural history model of HPV driven OPSCCs.

**Abbreviations:**

- HPV: Human Papillomavirus
- HNSCC: Head and Neck Squamous Cell Carcinoma
- OPSCC: Oropharyngeal Squamous Cell Carcinoma
BART: Bayesian Additive Regression Trees

NHANES: National Health and Nutrition Examination Survey

SEER: Surveillance Epidemiology and End Results Program

AUC: Area under the curve

ROC: Receiver operating characteristic

Author contributions: JOL, CM, PT, EK and ACP designed the study. EK and PT identified datasets and extracted data for model development. EK, ACP and CW developed the mathematical modelling approach. EK, PT and ACP drafted the manuscript. All authors edited, revised and approved the final manuscript for submission.

Funding: Health Research Board, Ireland. Grant/Award Numbers ICE-2015-1037.

Availability of data and materials: Code available at www.github.com/andrewcparnell/HNSCC. The SEER HPV data are restricted and are not included in the Github repository.

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Competing interests: Not applicable

Acknowledgements

We are grateful to the SEER HPV database for providing access to the restricted HPV data.

EK and ACP were paid for a portion of this work through HRB award ICE-2015-1037.

ACP is funded by Science Foundation Ireland Career Development Award grant 17/CDA/4695 and a SFI centre grant 12/RC/2289_P2
References


