MedCrave
Step into the World of Research
Biometrics & Biostatistics International Journal

Research Article
Open Access
CrossMark

Risk factor determinants and comparison of supervised machine learning algorithms to predict head and neck cancer

Abstract

Head and Neck Cancer (HNC) has emerged as a major public health concern in India. The Indian state Bihar, which ranks fourth in the total number of new cancer cases, head and neck cancer is the second most common cancer and has an increasing trend. The application of machine learning (ML) in disease diagnosis is increasing gradually. With this background, an attempt has been made to determine the risk factors and compare the performance of different variants of supervised ML algorithms for HNC prediction. The study confirms that poor oral hygiene, tobacco, alcohol and human papilloma virus (HPV) infections are the significant risk factor for HNC occurrence in Bihar. In comparison to all the variants of supervised machine learning algorithm, Random Forest showed maximum accuracy. This study will beneficial in medical decision support systems.

Keywords: head and neck cancer, Bihar, machine learning, risk factor, prediction

Volume 13 Issue 1 - 2024

Alok,[1] Rama Shanker,[2] Manoj Kumar Rastogi,[1] Ajay Vidyarthi,[3] Arun Kumar[3]
[1]Department of Statistics, Patna University, Patna, India
[2]Department of Statistics, Assam University, Silchar, India
[3]Mahavir Cancer Sansthan & Research Centre, Patna, India

Correspondence: Rama Shanker, Department of Statistics, Assam University, Silchar, India, Email shankerrama2009@gmail.com

Received: December 20, 2023 | **Published:** January 05, 2024

Introduction

In present days, cancer is a major burden of disease worldwide and has become a public health problem. It is a leading cause of death globally, accounting for an estimated 10 million deaths in 2020, or nearly one in six deaths among which 70% were from low - and - middle - income countries. Also in India, cancer incidence is continuing to rise. According to the Global Cancer Observatory (GLOBOCAN) 2020 report, India ranked third after China and USA in number of cases and it predicted that cancer cases in India would rise to 2.08 million with a rise of 57.5% in 2040 from 2020.[1] As per data of Union Ministry of Health and Family Welfare, among states, Uttar Pradesh topped with 2.10 lakh new cancer cases in 2022 followed by Maharashtra, West Bengal and Bihar, although the numbers are an underestimation of the actual figure because of under reporting of cancer cases due to shortage of hospital-based cancer registry (HBCR) and population-based cancer registry (PBCR)especially in Bihar. The global cancer burden continues to rise, putting enormous physical, emotional and financial strain on individuals, families, communities, and health-care systems.[2] Many health-care systems in low- and middle-income countries are inadequate to deal with this burden and a huge majority of cancer patients worldwide lack access to timely, high-quality diagnosis and treatment.[3] By avoiding risk factors and using currently available evidence-based prevention measures, between 30 and 50% of cancers are currently preventable.[4]

Head and Neck cancer is a general term that are classified by the anatomical areas such as lip, oral cavity, pharynx, larynx, nasal cavity, paranasal sinuses and salivary glands in which they arise.[5] Among all subsites, oral cavity cancer is more prevalent due to local custom of chewing betel nut with tobacco. Head and neck cancer is sixth most common cancer in the world.[6] Overall, 57.5% of worldwide HNC occur in Asian continent particularly in India, for both male and female, though males are affected significantly more than females.[7] In India, every year more than two lakh new HNC cases are diagnosed which constitutes 30% of all the cancers.[8] There has been an increase in the number of young adults affected by this disease.[9]

In Bihar, which is third most populated state and whose population density is highest among Indian states excluding union territory, head and neck cancer is the second most common cancer and continuously showing an increasing trend.[10] Here, nearly half of the population consumes tobacco, which is one of the major reasons behind rapid increase of HNC patients.[11] Being a backward and poor state, the health infrastructure is incapable to cater patients in early diagnosis and a result of it, more than 50% patients are diagnosed cancer in advanced stage.[8]

Head and Neck cancer is curable if detected in the early stages of the disease. Early diagnosis and prevention are better than cure. Therefore, the focus of the prevention strategy should be on raising public knowledge about risk factors because they can be changed. Cancer, as is frequently and rightly stated, is a disease that affects not only the patients but the entire family. So, considering the socio-economic impact of this disease, an attempt has been made to find out the risk factors associated with HNC in Bihar, where there is very much diversity in social customs, dietary habits, tobacco addiction and socio-economic status across the state.

In recent days, machine learning algorithm has shown a potential application in the area of disease prediction, which has recently gained significant attention from the data science research community.[12] This is due to the widespread adoption of computer-based technology in many forms in the health industry. A machine learning algorithm uses a variety of statistical, probabilistic and optimisation methods to learn from past data and identify useful patterns from large, unstructured and complex datasets. A supervised machine learning algorithm builds a model from a known set of input data (the learning set) and known responses to the data (the output) to make reasonable predictions for the response to new input data. So, in this paper, we have also compared the performance of different variant of supervised ML algorithm applied for head and neck cancer prediction which will help in improving the diagnosis accuracy and will beneficial in medical decision support systems. This study has been approved by ethics committee of ICMR- RMRIMS, Patna.

Material and methods

Study design

In the present study, to determine the risk factor associated with the head and neck cancer, a 1:2 unmatched case-control study design

has been selected. Cases are those in which the outcome or disease is already present and controls are those in which the disease/outcome is absent.

Sample size

By considering percentage of cases with exposure and percent of control exposed as 77.02 and 53.8 respectively estimated from the previous year hospital data and anticipated odds ratio as 2.87, power as 80%, two-sided confidence interval i.e. alpha as 5 %, and adjusting for 10% non-response rate, the minimum sample size for cases has been obtained. But for better precision and availability of resources, we have increased the sample size and recruited 100 cases and 200 controls since the ratio of control to cases is 2 for this study.

Source of data

To carry out the present study, data related to patients (Cases) have been collected from the HNC outpatient department (OPD) of the Mahavir Cancer Institute & Research Centre, Patna, which is a dedicated Cancer Hospital and one of the HBCR in the state. Only newly diagnosed cases of different anatomical subsites of Head and Neck cancer confirmed by histopathology have been included. Since head and neck cancer involves different anatomical subsites, so number of newly diagnosed cases of different subsites have been taken in proportion to patients diagnosed with different types during last two years in the same hospital.

Inclusion and exclusion criteria

For cases, inclusion criterion was: newly diagnosed HNC patient confirmed by histopathological and radiological tests and patients belonging to different districts of Bihar only. The exclusion criterion of cases was: patients coming for treatment from outside Bihar, malignancy proven patient referred from head and neck cancer OPD to other OPD and higher centre for treatment and patients coming for follow-up and recurrent disease treatment in OPD.

Similarly, only those respondents were chosen for control group who belonged to Bihar only and who came to head and neck cancer OPD for screening but were not diagnosed by histopathology or radiological tests and those attendants or relatives who came with diagnosed patients and whose age, socio-economic status, gender, habit and their standard of living were matched with patient up to a large extent. A written consent has been taken before including respondent either as case or control in the study.

Questionnaire

Data collection for the present study was carried out with the help of structured questionnaire which was divided into three parts. The first part covers the demographical profile of the respondent and it includes name, gender, address, cast, socio-economic status etc. Second part contains the information about the exposure and habit which leads the risk of developing head and neck malignancy such as occupation type, inhalation exposure, tobacco, alcohol, smoking habits, oral hygiene and HPV infection. The third part, only for case group, contains the disease information, where type of Head and neck cancer (anatomical subsite), date of diagnosis, method of diagnosis, disease stage, treatment type etc. are included.

Statistical methodology

To find the risk factor associated with occurrence of head and neck cancer, first Akaike information criterion (AIC) was used to find the best model among subset of models using optimal set of variables and to remove the multicollinearity, if existed. The model having lowest AIC value has been chosen as the best model. To measure the strength of association of the disease with exposure, odds ratio along with 95% confidence interval was evaluated using multiple logistic regression. The performance of different variants of supervised machine learning algorithms for predicting HNC has been compared using accuracy and other machine learning metric. All the analysis was carried out in RStudio statistical software package.

Results

Both case and control groups have almost same demographic characteristics. Male are predominant in numbers and the average age of both the groups are within (+/-) 5 years. Majority of the respondents in both the groups belong to the rural area with poor socio-economic background. For the present study *khaini,* betel nut and *gutka* has been combined into tobacco exposure and habit of *bidi* and cigarette is included in smoking exposure. It has been found that proportion of tobacco chewer is higher among cases than the controls. Though most respondents in both the groups have no inhalation exposure, but number of persons who are exposed to asbestos are highest in both the groups. The proportion of person who consume alcohol is more in case group as compared to control group. More than 90% HNC patients have poor oral hygiene. Having a family history of cancer is an established risk factor for the disease, but for the present study, maximum number of cases do not have family history of any type of malignancy Table 1 & 2.

Table 1 Demographic profile of cases and control group

Characteristics	Category	Cases (n)	Controls (n)
Gender	Male	82	156
	Female	18	44
Residence	Urban	32	79
	Rural	68	121
Religion	Hindu	91	161
	Muslim	9	39
	Others	0	0
Socio- Economic Status	Poor	76	93
	Middle	17	71
	Rich	7	36
Mean Age		49.97 years	45.47 years
Qualification	Illiterate	39	27
	Primary	21	35
	Middle	17	36
	10^{th}	11	33
	12^{th}	5	22
	Technical	6	24
	Graduate & above	3	20
Monthly Family Income	Less than 10,000	20	24
	10,000-25,000	56	76
	25,000-50,000	16	70
	50,000-1,00,000	1	6
	More than 1,00,000	7	24

Table 2 Exposure status of cases and controls

Exposure	Category	Cases	Controls
Occupational Inhalation	Wood Dust	5	7
	Asbestos	10	10
	Synthetic Fibers	9	5
	Radiation	3	6
	No Exposure	73	172
Alcohol	Yes	64	77
	No	36	123
Tobacco	Yes	91	104
	No	9	96
Smoking	Yes	28	37
	No	72	163
Dietary Habit	Vegetarian	18	56
	Non-Vegetarian	82	144
Family History	Yes	13	74
	No	87	126
HPV	Yes	27	23
	No	73	177
Oral Hygiene	Good	8	102
	Poor	92	98

In the present study, cases of all anatomical subsites of head and neck cancer have been included in proportion to patients diagnosed with different subsites during last two years in the same hospital and since oral cavity cancer is more prevalent in this area, therefore, accordingly their frequency was more in comparison to others in this study. More than half of the patients have been diagnosed in the advance stages of disease which is stage 3, 4A and 4B, where survival of patient is very poor. Most of the persons have taken treatment in this hospital after diagnosis but some patients moved to higher center or they denied treatment due to financial constraint and other reasons. In the treatment of HNC, radiotherapy plays a very significant role. So, number of patients who received radiation along with surgery was maximum. Compared to the number of patients who got neo - adjuvant therapy, a greater number of patients received adjuvant therapy Table 3.

Table 3 Disease information of patient studied

Particulars	Category	Frequency
Anatomical Subsite	Oropharyngeal	18
	Hypopharyngeal	6
	Laryngeal	10
	Nasopharyngeal	5
	Oral cavity	38
	Paranasal Sinus and Nasal cavity	5
	Salivary Gland	5
	Mets of Unknown Origin	6
	Lip	7
Stage	1	10
	2	21
	3	17
	4A	45
	4B	7
Received Treatment	Yes	84
	No	16
Type of Treatment Plan	Surgery (S)	8
	Chemotherapy (C)	5
	Radiotherapy (R)	17
	S+R	40
	C+R	4
	S+C	20
	S+C+R	6
Type of Therapy	Adjuvant	72
	Neo adjuvant	28

To find the risk factor associated with head and neck cancer, initially we have taken 13 independent variables which are supposed to be directly associated with the disease as per literature review and one dependent variable. The independent variable initially taken for analysis was gender, age, socio-economic status, residence, occupation, alcohol, tobacco, smoking, dietary habit, family history, oral hygiene, nutrition level and HPV. The dependent variable was Type which shows whether there is malignancy present or not. Now, among different subset of models, best model has been selected based on the lowest AIC value of 263.75 which contains only 9 predictors. Presence of multicollinearity among the predictors in the selected model has been checked using generalized variance inflation factor (GVIF) which is the square root of the VIF for individual predictor. Since the value of GVIF for all variables are around 1, so we can conclude that there is no multicollinearity present among predictors and all are independent.

Among all the considered risk factors, alcohol (OR=2.473, $p<0.05$), tobacco (OR=10.28, $p<0.01$), oral hygiene (OR=7.386, $p<0.01$), and HPV infection (OR=3.98, $p<0.05$) results to be significantly associated with the incidence of head and neck cancer. Although there are differences between the magnitudes of odds ratio as tobacco being the highest followed by oral hygiene, HPV and alcohol Table 4 & Figure 1.

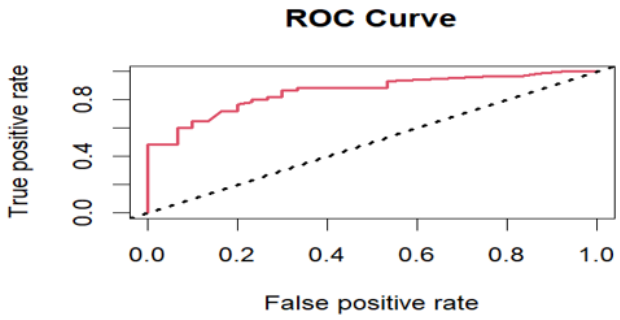

Figure 1 Receiver operating curve.

The value of AUC (Area Under Curve) results out to be 85.72% which shows that model's ability to distinguish between binary outcome of dependent variable correctly is very good. The confusion matrix, which determines the diagnostic ability of ML classification algorithm, obtained from logistic regression, is shown in Table 5.

The comparative performance of different supervised machine learning algorithm to predict head and neck cancer using accuracy, sensitivity, specificity and F1 score is shown in Table 6. Among all variants of ML algorithm, Random Forest showed the maximum accuracy followed by Decision tree and Naïve Bayes.

Table 4 Odds Ratio with 95 % confidence interval of different predictors associated with head and neck cancer

Predictors	Odds ratio	p-value	95% Confidence interval
Intercept	0.035	6.58e-05	0.006-0.167
Gender			
Male	0.374	0.108	0.109-1.234
Female	Ref*		
Socio-Economic Status			
Rich	0.321	0.139	0.065-1.371
Poor	1.273	0.631	0.472-3.450
Middle	Ref		
Alcohol			
Yes	2.473	0.045*	1.026-6.132
No	Ref		
Tobacco			
Yes	10.288	0.0001*	3.378-38.661
No	Ref		
Dietary Habit			
Vegetarian	0.439	0.097	0.159-1.134
Non- vegetarian	Ref		
Family History			
Yes	0.445	0.122	0.152-1.211
No	Ref		
Oral Hygiene			
Poor	7.386	0.0001*	2.831-21.868
Good	Ref		
Nutrition Level			
Sufficient	0.557	0.202	0.222-1.362
Insufficient	Ref		
HPV			
Yes	3.986	0.014*	1.380-12.812
No	Ref		

*Ref denotes reference category

Table 5 Confusion matrix obtained from logistic regression algorithm

	Actual	
Predicted	**Yes**	**No**
Yes	21	9
No	9	51

Table 6 Comparison of different supervised machine learning algorithm

Supervised ML Algorithm	Accuracy (%)	Sensitivity	Specificity	F1 Score
Logistic Regression	80	0.70	0.85	0.70
Decision tree	81.1	0.70	0.87	0.71
Random Forest	82.2	0.73	0.86	0.73
Naïve Bayes	80	0.7	0.85	0.7

Discussion

As per population-based cancer registries, head and neck cancer is the commonest cancer in Indian men and third most common in women due to widespread use of tobacco. According to Global Adult Tobacco Survey, in India, the tobacco consumption rate among adults is 34.6% which is higher in males compared to females and more prevalent in rural areas.[13] Among all the subsite, oral cancers are most common which constitute about 40 % of all subsites. The reason for high mortality due to this disease in India includes diagnosis at advanced stage, lack of medical facility, poor oral nutritional status and other socio-economic factors. Bihar, which is a poor and backward state, the health infrastructure is incapable to cater patients in early diagnosis and due to this reason; more than 50% patients are diagnosed in advanced stage of disease. Cancer is a disease that impacts not only the patient but the entire family is suffered mentally, physically and financially as well. So, considering the socio-economic impact of illness, it is critical to recognize the patterns and important risk- factor responsible for it. In recent times, machine learning has shown an immense potential in improving accuracy and speed of cancer diagnosis.[14] Improving patient outcomes and generating previously unattainable medical insights are also the main objectives of machine learning.[15] In supervised variant of machine learning, a prediction model is developed using labelled training dataset and then this model is fed on the unlabeled test dataset to predict the labels. So, in this paper, which is basically a hospital based retrospective study, we have tried to find out the risk factor responsible for HNC in Bihar using ML technique. An attempt has also been done to assess the performance of different variant of supervised machine learning algorithm in predicting head and neck cancer and results from various algorithms have been compared among themselves. Akaike Information Criterion has been used to find the best model that explains most variation in data, while penalizing for models that use an excessive number of parameters. AIC is similar to adjusted R- squared because it also penalizes for adding more variables to the model. The absolute value of AIC does not have any significance and is used for comparison purpose. So, we only compare AIC value whether it is increasing or decreasing by adding more variables in the model and in the case of multiple models, the one which has lower AIC value is preferred. The model having lowest AIC value of 263.75 has been chosen as the best model. By using multiple logistic regression, we find that among all the predictors, alcohol (OR=2.47, CI=1.026-6.132), tobacco (OR=10.288, CI= 3.378-38.661), oral hygiene (OR=7.386, CI=2.831-21.868) and HPV infection (OR= 3.986, CI =1.380-12.812) are significantly associated with the incidence of head and neck cancer. The above study also confirms Human Papilloma Virus (HPV) as an important risk factor of HNC. While HPV is generally positive in the patients of oropharyngeal cancer but in this study, it has been tested in patients of all anatomical subsites. ROC is a performance measurement metric of a classification model and is created by plotting the true positive rate against the false positive rate at different threshold values. Area under the Curve (AUC) represents the area under the ROC curve and it measures the overall performance of the binary classification model. The AUC measurement value for the above model results to 85.72% which reveals model performance is very good in prediction.

Again, we have used another variant of machine learning algorithms such as Decision tree, Random Forest and Naïve Bayes to build a model that accurately predicts head and neck malignancy and calculated accuracy, sensitivity, specificity and F1 score obtained from them. Random forest which combines the output of multiple decision trees to reach a single result gives the maximum accuracy followed by decision tree and logistic regression. The metric of naïve bayes and logistic regression are found to be equal.

Although this study was hospital based but it highlights the risk factor profile of HNC for whole state since this hospital caters patients from every part of state. It helps in understanding the possible risk factor and behaviour patterns of HNC patients. This study recommends an urgent need for taking appropriate preventive strategies through common risk factor approach such as ensuring tobacco ban in the state along with screening programme for early

detection. With the advent of new ML techniques, there is no need to depend upon only one algorithm to predict disease. So, in this study four different classification algorithm have been used to build model for better prediction and their performance have been compared using metric obtained from them.

Conclusion

Head and Neck Cancer is the second most prevalent cancer in Bihar after breast cancer.[10] Since a greater number of patients are diagnosed in advanced stage of disease, their survival is also poor. The results obtained from the above epidemiological study on the population of Bihar confirms poor oral hygiene, tobacco consumption, alcohol intake and HPV infection as the significant risk factor for head and neck cancer burden in the state. Among all the subsite of HNC, oral cavity cancer is predominant and maximum patients present with poor oral hygiene. Males are mostly affected with HNC disease and in almost all, the habit of tobacco consumption was present. The scope of this research paper incorporates the comparison of the performance of different variant of supervised machine learning algorithm applied for head and neck cancer prediction using different metric. The important result of performance comparison can be used to help researchers in the selection of appropriate supervised machine learning algorithm for their studies.

Acknowledgments

Authors are thankful and would like to acknowledge the contributions of the doctors and staffs of the Head and Neck Cancer Department of Mahavir Cancer Sansthan and Research Centre, Patna, Bihar, India. Authors are also grateful to the editor in chief and the anonymous reviewer of the journal for comments which improved the quality of the paper.

Conflicts of interest

None.

Funding

None.

References

1. Sathishkumar K, Chaturvedi M, Das P, et al. Cancer incidence estimates for 2022 & projection for 2025: result from National Cancer Registry Programme, India. *Indian J Med Res*. 2022;156(4&5):598–607.
2. World Health Organization (WHO). WHO report on cancer: setting priorities, investing wisely and providing care for all. Geneva, Switzerland: 2020.
3. Chalkidou K, Marquez P, Dhillon PK et al. Evidence-informed frameworks for cost-effective cancer care and prevention in low, middle, and high-income countries. *Lancet Oncol*. 2014;15(3):e119–131.
4. Kerschbaum E, Nüssler V. Cancer prevention with nutrition and lifestyle. *Visc Med.* 2019;35(4):204–209.
5. Argiris A, Karamouzis MV, Raben D, et al. Head and neck cancer. *Lancet*. 2008;371(9625):1695–1709.
6. Vigneswaran N, Williams MD. Epidemiologic trends in head and neck cancer and aids in diagnosis. *Oral and maxillofacial surgery clinics of North America*. 2014;26(2):123–141.
7. Kulkarni MR. Head and neck cancer burden in India. *Int J Head Neck Surg* 2013;4(1):29–35.
8. Chauhan R, Trivedi V, Rani R, et al. A study of head and neck cancer patients with reference to tobacco use, gender, and subsite distribution. *South Asian J Cancer.* 2022;11(1):46–51.
9. Kanmodi K, Nnebedum N, Bello M, et al. Head and neck cancer awareness: a survey of young people in international communities. *Int J Adolesc Med Health*. 2019;33(4):2018–0231.
10. Pandey A, Raj S, Madhawi R, et al. Cancer trends in Eastern India: retrospective hospital-based cancer registry data analysis. *South Asian J Cancer*. 2019;8(4):215–217.
11. Singh R, Alok. Recent changes in health status of women in bihar through national family health survey window. *J Clin of Diagn Res*. 2018;12(4):IE01–IE05.
12. Uddin S, Khan A, Hossain ME, et al. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak*. 2019;19(1):281.
13. Dandekar M, Tuljapurkar V, Dhar H, et al. Head and neck cancers in India. *J Surg Oncol*. 2017;115(5):555–563.
14. Kourou K, Exarchos TP, Exarchos KP, et al. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2014;13:8–17.
15. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*. 2019;19(1):64.