International Journal of
Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2025; 10(12): 95-101
© 2025 Stats & Maths
https://www.mathsjournal.com
Received: 02-10-2025
Accepted: 05-11-2025

Jayanta Sarma Kakoty
Research Scholar, Department of Mathematics, University of Science & Technology, Meghalaya, India

Parbin Sultana
Professor, School of Technology and Management, University of Science & Technology, Meghalaya, India

Corresponding Author:
Jayanta Sarma Kakoty
Research Scholar, Department of Mathematics, University of Science & Technology, Meghalaya, India

Factors affecting oral cancer prognosis in Assam: A binary logistic regression study

Jayanta Sarma Kakoty and Parbin Sultana

DOI: https://www.doi.org/10.22271/maths.2025.v10.i12b.2212

Abstract

Oral cancer is a major public health issue in Assam, largely due to the prevalent use of areca nut, tobacco, smoking and alcohol. This study uses binary logistic regression to examine factors influencing the prognosis of 1000 oral cancer patients. Patient status (either “Death” or “Alive”) is the dependent variable, with factors categorized as Personal (age, gender, behavioural factors, education, family income, food habits and locality [rural / urban]) and Clinical (cancer-directed treatment [CDT], follow-up, topography sites, stage groups, treatment given prior to registration [TGPR], metabolic risk factors [MRF] [over weight, raised BP and raised BG]). The model correctly predicts 99.0% when selected variables are included. Significant factors include Behavioural factors, Family Income, CDT, Follow up, Stage groups, TGPR, MRF and Locality (Rural / Urban). The non-significant factors include age, gender, education, food habits and topography sites. The study concludes that more consumption of alcohol, not assigned CDT, TGPR, higher MRF and the patients of urban areas increase the risk of oral cancer and the others have no significant impact. The model fits the data and demonstrates strong predictive accuracy.

Keywords: Rate of metabolism, blood mass stream rate, warm conductivity, warm era, limited component method, Pennes Bio - Heat Model

Introduction

Cancer is a disease caused by an uncontrolled division of abnormal cells in the parts of the body. Cancer has become more than just a global health problem and is also a cause of deep suffering to the individual patients, their families and to the community at large. GLOBOCAN (Global Cancer Observatory), a project of the International Agency for Research on Cancer (IARC) provides data on Cancer incidence, mortality and geographic variability. The data is used to predict the future magnitude of cancer incidence and to evaluate national cancer control efforts. The GLOBOCAN 2020 estimates that there were 19.3 million new cases of cancer and almost 10 million deaths from cancer in 2020. According to World Health Organisation (WHO), cancer is found to be second leading cause of death throughout the global *[1]*. National Centre for Disease Informatics and Research (NCDIR) situated in Begaluru was established in 2011 by the Indian Council of Medical Research (ICMR). However, the ICMR-NCDIR coordinates the National Cancer Registry Programme (NCRP), which was established in 1981. The ICMR-NCDIR provides training, technical support, data evaluation and analysis to the Population Base Cancer Registries (PBCR). According to ICMR-NCDIR latest report *[2]*, oral cancer is among the top three cancers in India and it is among the top five cancers in Assam. The study is trying to analyse the prognosis of oral cancer in different categories of patients in the region. In this study, a logistic regression model will be applied to understand the factors influencing the prognosis of the oral cancer in the region. In this study the probable factors influencing the prognosis of oral cancer have been classified into two categories, viz., personal factors and clinical factors. The study is based on the patients database (2015 - 2020) is provided by North East Cancer Hospital & Research Institute (NECHRI), which established in 2008 has emerged as the most comprehensive and leading cancer hospital in North East India located in Jorabat, G.S. Road, Guwahati-781023. Its therapeutic and diagnostic facilities, according to several eminent experts are comparable with national level hospitals. The reason of choosing NECHRI for the study area is that it is one of the best premier cancer research institutes in India, the cancer patients of all the districts of Assam come to the institute to avail their treatments.

Personal Factors

Personal factors include Demographic factors, Behavioural factors, socioeconomic factors and Food habits. Demographic factors are classified into age and gender. Behavioural factors are the factors that the patients are habituated to abuse substances like areca nut, tobacco, smoking and alcohol. In Socioeconomic factors, it is considered as Education and Family income (pa). Education is classified as under metric, under graduate and graduate or others. Food habits include whether the patients taking foods veg or non-veg.

Clinical Factors

Clinical factors include Cancer directed treatment (CDT), Follow up (in years), Topography sites, Stage groups, Treatment given prior to registration (TGPR) and Metabolic risk factors (MRF).

Cancer Directed Treatment (CDT)

Cancer directed treatment (CDT) has five options - Yes, No, Treatment advised but not accepted (TANA), and Unknown.

Follow Up (FU)

Follow up of a patient is determined by the difference between Date of First Diagnosis (DFD) and Date of death (DOD) in case of the patients who are death or Date of last contact (DOL) in case of the patients who are alive till the date of interview (DOI).

Topography Sites of Oral Cancer (TSOC)

In this study, Topography sites of oral cancer found in different patients, are classified as follows:

- **Cheek:** Buccal Mucosa (BM), Buccal Sulcus (BS), Cheek Mucosa (CM)
- **Esophagus / Throat / Nose:** Aryepiglottic Fold (AF), Hypopharynx (Hy), Hypopharyngeal Wall (HW), Lateral Wall Pharynx (LWP), Lateral Wall of Nasopharynx (LWN), Laryngopharynx (La), Nasopharynx (Na), Nasopharyngeal Wall (NW), Oropharynx (Op), Posterior Wall of Oropharynx (PWO), Posterior Wall of Pharynx (PPW/PWP), Posterior Wall of Hypopharynx (PWH/PHW), Pyropharynx Sinus (PS/PFS), Submandibular Gland (SG), Tonsillar Fossa (TF), Tonsillar Pillar (TP), Tonsil (Ts), Vallecula (Va).
- **Gingiva:** Lower Gingiva (LG), Maxillary Gingiva (MG), Gingiva Buccal Sulcus (GBS)
- **Lip:** Lower Lip (LL), Mucosa of Lip (MOL/MUO)
- **Mouth:** Floor of Mouth (FOM), Oral Cavity (OC), Retromolar Trigon (RT/RMT), Parotid Gland (PG).
- **Palate:** Hard Palate (HP), Soft Palate (SP)
- **Tongue:** Border of Tongue (BOT), Dorsal Surface of Base of Tongue (DSBT)

Stage Groups (SG)

American Joint Committee on Cancer (AJCC)*[14]* defines the different categories of oral cancer by TNM (Tumor, Node, Metastasis) principle as follows:

Table 1: Definition of Primary Tumor (T)

T Category	T Criteria
T_x	Primary tumor cannot be assessed
T_{is}	Carcinoma in situ
T_1	Tumor ≤ 2 cm, ≤ 5 mm depth of invasion (DOI)
T_2	Tumor ≤ 2 cm, DOI > 5 mm and ≤ 10 mm OR Tumor > 2 cm but ≤ 4 cm and ≤ 10 mm DOI
T_3	Tumor > 4 cm OR any tumor > 10 mm DOI
T_4	Moderately advanced local disease (T_{4a}), Advanced local disease (T_4_b)

Table 2: Definition of Regional Lymph Node (N) Clinical N (cN)

N Category	N Criteria
N_x	Regional lymph nodes cannot be assessed
N_0	No regional lymph node metastasis
N_1	Metastasis in a single ipsilateral lymph node, 3 cm or smaller in greatest dimension, ENE(-)
N_2	Metastasis in a single ipsilateral node >3 cm but ≤6 cm, ENE(-) OR metastases in multiple ipsilateral nodes, none >6 cm, ENE(-) OR bilateral/contralateral nodes, none >6 cm, ENE(-)
N_{2a}	Metastasis in a single ipsilateral node >3 cm but ≤6 cm in greatest dimension, ENE(-)
N_2_b	Metastasis in multiple ipsilateral lymph nodes, none >6 cm in greatest dimension, ENE(-)
N_2_c	Metastasis in bilateral or contralateral lymph nodes, none >6 cm in greatest dimension, ENE(-)
N_3	Metastasis in a lymph node >6 cm, ENE(-) OR metastasis in any node(s) with clinically overt ENE(+)
N_{3a}	Metastasis in a lymph node >6 cm in greatest dimension, ENE(-)
N_3_b	Metastasis in any node(s) with clinically overt ENE(+)

Table 3: Definition of Distant Metastasis (M)

M Category	M Criteria
M_x	Distant metastasis cannot be assessed
M_0	No distant metastasis
M_1	Distant metastasis

The prognostic stage groups are shown in the table 4.

Table 4: Prognosis stage groups of oral cancer

When T is	And N is	And M is	Then the stage group is
T1	N0	M0	I
T2	N0	M0	II
T3	N0	M0	III
T1,2,3	N1	M0	III
T4a	N0,1	M0	IVA
T1,2,3,4a	N2	M0	IVA
Any T	N3	M0	IVB
T4b	Any N	M0	IVB
Any T	Any N	M1	IVC
Not Detected	Not Detected	Not Detected	0

Source: AJCC (American Joint Committee on Cancer), 2023

Treatment given prior to registration (TGPR)

TGPR has three options - Yes, No and Unknown.

Metabolic Risk Factors (MRF)

Metabolic syndrome is not a disease in itself. Instead, it's a group of risks factors- high blood pressure, high blood sugar, unhealthy cholesterol levels and abdominal fats. Occurrence of these factors in a person are influenced by his / her lifestyle. The metabolic factors generally support the growth of any disease. According to NCDIR *[2]* report, it is found that 80% of cancer patients have the following metabolic factors-

- Overweight / Obese BMI > 25
- Raised blood pressure
- Raised blood glucose (random)

The study is trying to analyse the two probable factors - Personal factors and Clinical factors influencing the prognosis of Oral cancer in context of Assam. In this analysis, a logistic regression model, have been applied to study the influence of these two factors on the status of the patients in term of Death or Alive.

Review of Literature

Different literatures on related topics have been reviewed and some of them are present as follows.

- Babu B. Madana Mahana (2008)*[4]* discussed the various treatments and compare the effects of the treatments by observing the responses of the cancer patients by using various statistical tools basically probability distributions and logistic regression models. By the study, it was observed that male patients have more risk than female patients, the risk in discontinue treatments of the patients is higher than those patients who are continuing the treatments, the patients who are in stage 4 are having higher risk when compared to the patients who are in stage 2 and the male patients who are above 80 years are having more risk when compared to other age groups.
- Hamidi A.Al-Hakimi *et al.* (2016)*[5]* mentioned that knowledge of Oral Cancer affects early detection and diagnosis of this disease. The study aimed to assess the current level of public knowledge of oral cancer and examine how demographic background factors affect this knowledge. A logistic regression model was utilized with demographic background variables as independent variables and knowledge of oral cancer as dependent variable. A path analysis was conducted to build a structured model. The outcomes of the study were that the maximum participants associated with the study had no knowledge about the oral cancer. The model interpreted that age, place of residence and education levels were significantly associated with knowledge levels.
- Li-Chen Hung *et al.*(2020)*[6]* performed a univariate Poisson Regression analysis and the incidence density ratio (IDR) was used to indicate the annual risk of oral cancer incidence per 100,000 persons. A logistic regression model was used as a predictive model. The results showed that the oral cancer incidence rate was significantly higher among betel nut chewers than among smokers and significantly higher among individuals with concurrent habits of smoking and betel nut chewing than among individuals with either habit in Taiwan.
- Schwartz Stephen M. (1998)*[7]* used standard methods for the statistical analysis of case-control studies. The odds ratio (OR) for the association between oral cancer risk and factors under study was estimated using unconditional logistic regression. By the study, it was observed that among the males only, oral squamous cell carcinoma (SCC) risk increased with self reported decreasing age at first sexual activity, increasing the number of sex partners, cigarette smoking and alcohol consumption.
- Abedin Tasnima, Chawdhury Mohammad Zia et. al. (2016)*[13]* applied the logistic regression model to analyse how the subject's age and smoking status is related to his / her cardiovascular disease (CVD) status. The results support our hypothesis about the hypothetical data that subject's age and smoking status is related to his/her CVD status. The likelihood of a subject's having CVD is positively related to his/her age. However given the same age, smokers are more likely to have CVD than nonsmokers.
- Zangmo Choney, Tiensuwan Montip (2018)*[12]* concluded that the logistic regression models that can identify the factors which affect the status of last contact of the cancer patients, are also used predict the status of cancer patients. The results showed that death occurred more to male cancer patients than female patients. The death percentage of the male patients was higher than the female patients although females were the ones who were mostly diagnosed with cancers. For the cancer site, most of the male patients suffered from digestive organ cancer. For female, the genital organs cancer had the highest number of occurrences. There are few numbers of male patients who were diagnosed with genital organs cancer.
- Suresh Neena, Thomas Binu, Joseph Jeena (2024)*[15]* aims to investigate the research productivity of heart disease classification using logistic regression model to analyse the current patterns and potential future trends through bibliometric analysis. The study finds that upon arrival of a new patient, during the admission time itself, the regression model will enable us to identify the risk factors associated with the patients and make appropriate decision - a patient under high risk could immediately undergo further diagnostic tests or receive proper care.
- Huzain Aziz (2024) demonstrated the efficacy of Logistic Regression in predicting heart disease using a well - structured dataset, achieving considerable accuracy and high recall across a 5-fold cross-validation framework. Results indicated that Logistic Regression performs robustly, with accuracy ranging from 80% to 88.29%,

and high recall rates, highlighting its potential as a valuable tool in medical diagnostics.

- Izmy Alwaiah *et al* (2024) assesses the predictive power of Logistic Regression in forecasting liver disease prevalence within the Indian demographic, specifically leveraging a dataset of the patient records from the North East of Andhra Pradesh. The findings reveal moderate to high accuracy (69.23% to 74.14%) and precision (59.62% to 75.62%) levels, indicating Logistic Regression potential as a viable predictive tool in predicting liver disease in medical diagnostic.
- Ahmed Shaker Abdalrada *et al* (2024) proposes a logistic regression model to anticipate the likelihood of Diabetes Syndrome incidence. By using the Pima Indians Diabetes dataset, the model predicted accuracy rate of 77.6% with a sensitivity of 72.4%, specificity of 79.6%, Type-I error of 27.6% and Type-II error of 20.4%.

Research Questions

The Research questions related to the report are given below:

- What the various personal and clinical data are required to study the prognosis of oral cancer in context of Assam.
- How the personal and clinical factors are associated with the status of the patient.
- Why the model is applied to study the prognosis of oral cancer in the region.
- How the model is used for predictions.

Research Methodology

The selected sample is based on the patients database provided by North East Cancer Hospital and Research Institute (NECHRI) during the period 2015 - 2020. The dependent variable $\hat{Y}$ used in this study is the number of deaths or lives. The factors which influence the status of the patients (response variable) are considered as predictor variables. The categorial values of the variables associated with Personal data and Clinical data of the patients are given in the Table 5

Table 5: The categorial values of the variables associated with Personal data and Clinical data of the patients

Personal Factors (X_1-X_6)	Categorical Values	Clinical Factors (X_7-X_{13}, Y)	Categorical Values
Demographic Factors		CDT (X_7)	1 = Yes 2 = No 3 = Treatment assigned but not detected
Age in years (X_1)	Positive integer	Follow-up (X_8)	Positive integer
Gender (X_2)	1 = Female 2 = Male	Topography sites (X_9)	1 = Cheek 2 = Esophagus / Throat / Nose 3 = Gingiva 4 = Lip 5 = Mouth 6 = Palate 7 = Tongue
Behavioural Factors (X_3)	1 = Consumes nothing 2 = Areca nut 3 = Tobacco 4 = Smoking 5 = Alcohol 6 = Areca nut + Tobacco 7 = Areca nut + Smoking 8 = Areca nut + Alcohol 9 = Tobacco + Smoking 10 = Tobacco + Alcohol 11 = Alcohol + Smoking 12 = Areca nut + Tobacco + Smoking 13 = Areca nut + Tobacco + Alcohol 14 = Areca nut + Smoking + Alcohol 15 = Tobacco + Smoking + Alcohol 16 = All habits	Stage groups (X_{10})	1 = Unknown / Not detected 2 = Stage 1 3 = Stage 2 4 = Stage 3 5 = Stage 4
Education (X_4)	1 = Under-Metric (UM) 2 = 10th Passed to Undergraduate (UG) 3 = Graduate or Higher (GH)	TGPR (X_{11})	1 = Yes 2 = No
Family income (₹/year) (X_5)	1 = Below 1 lakh 2 = 1-5 lakhs 3 = 6 lakhs and above	MRF (X_{12})	1 = None 2 = Overweight 3 = Raised BG 4 = Raised BP 5 = Overweight + Raised BG 6 = Overweight + Raised BP 7 = Raised BG + Raised BP 8 = Overweight + Raised BG + Raised BP
Food habits (X_6)	1 = Vegetarian (V) 2 = Non-vegetarian (NV)	Rural / Urban (X_{13})	1 = Rural 2 = Urban

		Dependent Variable (Y)	Y = 0 → Death
			Y = 1 → Alive

The statistical model used here is Binary Logistic Regression Model, where, the dependent variable (Y) is defined by the patient status.

$$Y = \begin{cases} 0, \mathit{if\ the\ patient\ is\ death} \\ 1, \mathit{if\ the\ patient\ is\ alive} \end{cases}$$

Let, $\Pr(Y = 1) = p$

$\therefore \Pr(Y = 0) = 1 - p$

$where, 0 < p < 1$

The Logit is defined as the logarithm of the odds ratio $\frac{p}{1-p}$, which is given by the equation

$$Logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i X_i$$

The fitted value is $p = \frac{e^{Logit(p)}}{1+e^{Logit(p)}}$

To test the fitness of the model Omnibus tests or Hosmer Lemeshow tests are applied which are based on the statistic chi-square.

For clinical data, the following additional terms are considered.

- Date of First Diagnosis (DFD)
- Date of Last Contact (DOL)
- Date of Death (DOD)
- Date of Interview (DOI)

Follow Up by a patient is determined as follows

$$= \begin{cases} \frac{DOD - DOF}{365}, if\ the\ patient\ was\ death\ before\ the\ interview \\ \frac{DOL - DOF}{365}, if\ the\ patient\ was\ alive\ till\ the\ interview \end{cases}$$

Analysis and Interpretations

The effective sample size taken for this study is 1000 patients all over the Assam. The questionnaire is designed from the proforma given by Indian Council of Medical Research (ICMR) (Annexure1). Personal data of the patients are collected by taking interview of the patients or relatives over phone and personal interviews were also conducted who are accessible and the clinical data of the patients are collected from the hospital records.

Analysis of the model

The data are analysed by applying logistic regression model in SPSS. The outcomes are given below.

Table 6: Case Processing Summary

Unweighted Cases[a]		**N**	**Percent**
Selected Cases	Included in Analysis	1000	100.0
	Missing Cases	0	0
	Total	1000	100.0
Unselected Cases		0	0.0
Total		1000	100.0

Table 7: Dependent Variable Encoding

Original Value	**Internal Value**
Death	0
Alive	1

By the Table 6, it is observed that all the cases are included in the analysis and so there is no missing case. Patient status is the dependent binary variable taking 0 as the patient is death and 1 as the patient is alive as shown in the Table 7.

Table 8: Classification Table[a]

	Observed		**Predicted**		
			Status		**Percentage Correct**
			Death	**Alive**	
Step 1	Status	Death	731	5	99.3
		Alive	5	259	98.1
	Overall Percentage				99.0
a. The cut value is 0.500					

By the Table 8, out of 1000 cases, the number of true death is 731, the number of false death is 5, the number of true alive is 259 and the number of false alive is 5. The overall true percentage is $\frac{731+259}{1000} \times 100\% = 99.0\%$. Hence, 99.0% of the data of dependent variable (patient status) are correctly explained by the model.

To test the significance of the model, Omnibus Tests are applied and are shown in Table 9

Table 9: Omnibus Tests of Model Coefficients

		Chi-square	**df**	**Sig.**
Step 1	Step	1096.903	13	.000
	Block	1096.903	13	.000
	Model	1096.903	13	.000

From Table 9, it is observed that p-values are < 0.05, i.e. the model coefficients are significant. It indicates that the inclusion of the explanatory variables has significantly improved the model's to explain the response variable Y.

Table 10: Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1.736	8	0.988

From Table 10, it is observed that p-value > 0.05 which indicates a good fit, meaning the model's predictions align well with the actual data.

Table 11: Significance of the coefficients of the model

		B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Step 0	Constant	-1.025	0.072	204.253	1	0.000	0.359

Table 12: Test of significance of variables (factors)

		β	**S.E.**(β)	**Wald**	**df**	**Sig.**	**Exp** (β)
Step 1[a]	Age in years	-0.019	0.030	0.397	1	0.529	0.981
	Gender	2.071	11.746	0.031	1	0.860	7.930
	Behavioural Factors	-0.377	0.126	8.966	1	0.003	0.686
	Education	-0.052	0.662	0.006	1	0.937	0.949
	Family Income	2.393	0.961	6.202	1	0.013	10.952
	Food habits	-1.089	1.503	0.525	1	0.469	0.337
	CDT	-1.437	0.555	6.687	1	0.010	0.238
	Follow up	2.104	0.477	19.434	1	0.000	8.195
	Tophography sites	-0.055	0.176	0.099	1	0.753	0.946
	Stage groups	-2.166	0.407	28.361	1	0.000	0.115
	TGPR	-3.711	1.054	12.403	1	0.000	0.024
	MRF	-2.286	0.477	22.957	1	0.000	0.102
	Locality (Rural / Urban)	-2.588	0.894	8.378	1	0.004	0.075
	Constant	10.997	23.814	0.213	1	0.644	59714.579
a. Variable(s) entered on step 1: Age in years, Gender, Behavioural Factors, Educational Qua, Family Income, Food habits, CDT, Follow up, Tophography sites, Stage groups, TGPR, MRF, Rural / Urban.							

By the Table 11, it may be concluded that the coefficients are$\neq 0$, as the p-value is less than 0.05. It is seen that the intercept - only model is ln (odds) = -1.025. It is found that the predicted odds in favour of 'Alive' is Exp (β) = 0.359.

By the Table 12, it is observed that the variables Behavioural factors, Family Income, CDT (Cancer Directed Treatment), Follow up, Stage groups, TGPR (Treatment given prior to registration), MRF (Metabolic Risk Factors) and Loocality (Rural / Urban) are significant as their p-values are less than 0.05 when the other variables are not included. In other words, the dependent variable the 'Status' (Alive) is more influenced by these variables. The remaining variables (or factors) are not significant as their p-values are > 0.05. It implies that the response variable Y, the status of the patient is very less influenced by the factors Age, Gender, Education, Food habits and Topography sites

The coefficient β is associated with the respective variable (factor) and Exp(β) indicates the odds in favour of 'Alive'. The coefficients of the model of the significant factors (variables) interpret that

- The coefficient of the behavioural factor is negative and $E(\beta) = 0.686 < 1$, it implies that the rate of lives is 0.686 times lower, if the behavioural factor is increased by one unit. In other words, the mortality rate increases with the increase of behavioural factors.
- The coefficient of the family income is positive and $E(\beta) = 10.952 > 1$, it implies that the rate of lives is 10.952 times higher, if the family income is increased by one unit. In other words, the mortality rate is lower in higher family income groups.
- The coefficient of the CDT (Cancer Directed Treatment) is negative and $E(\beta) = 0.238 < 1$, it implies that the rate of lives is 0.238 times lower, if the CDT is increased by one unit. In other words, the mortality rate increases when CDT is not assigned or CDT assigned but not detected.
- The coefficient of the follow up time is positive and $E(\beta) = 8.195 > 1$, it implies that the rate of lives is 8.195 times higher, if the follow up time is increased by one unit. In other words, the mortality rate decreases with the increase of the follow up time.
- The coefficient of the stage group is negative and $E(\beta) = 0.115 < 1$, it implies that the rate of lives is 0.115 times lower, if the stage group is increased by one unit. In other words, the mortality rate increases in higher stage groups.
- The coefficient of the TGPR (Treatment given prior to registration) is negative and $E(\beta) = 0.024 < 1$, it implies that the rate of lives is 0.024 times lower, if the TGPR is increased by one unit. In other words, the mortality rate quite increases if the treatment is not given prior to registration.
- The coefficient of the MRF (Metabolic Risk Factors) is negative and $E(\beta) = 0.102 < 1$, it implies that the rate of lives is 0.102 times lower, if the MRF is increased by one unit. In other words, the mortality rate increases in higher metabolic risk factors.
- The coefficient of the locality (rural / urban)) is negative and $E(\beta) = 0.075 < 1$, it implies that the rate of lives is 0.075 times lower, if the locality is increased by one unit. In other words, the mortality rate is higher in urban areas.

The outliers of the response variable Y (the status of the patients) by using the model are shown in Table 13

Table 13: Outliers of the response variable

Case	Selected Status[a]	Observed	Predicted	Predicted Group
		Status		
125	S	D**	0.854	A
380	S	D**	0.868	A
473	S	D**	0.852	A
534	S	A**	0.008	D
537	S	A**	0.165	D
544	S	A**	0.099	D
594	S	D**	0.844	A
775	S	A	0.644	A

a. S = Selected, U = Unselected Cases and ** = Misclassified Cases

From Table10, it is observed that there are eight outliers. Among them the number of misclassified (false predicted) cases is seven and there is one true prediction.

Conclusion

Binary logistic regression is performed to study the probable factors influencing the prognosis of oral cancer in the state of Assam. The influenced factors are classified as Personal factors and Clinical factors. The personal factors contain Age, Gender, Behavioural factors - Areca nut, Tobacco, Smoking, Alcohol, Education, Family income, Food habits and locality (rural / urban). The clinical factors contains Cancer directed treatment (CDT), Follow up, Topography sites, Stage groups, Treatment given per registration (TGPR) and Metabolic risks factors (MRF) - Overweight, Raised BP and Raised BG. The total number of cases (patients) studied here is 1000. The dependent variable is the patient status which is either "Death" or "Alive".

Based on the study, the following conclusions can be drawn

- 99.0% of the data of dependent variable (patient status) are correctly explained by the model.
- the inclusion of the explanatory variables has significantly improved the model's to explain the response variable Y, the status of the patients (By Omnibus tests)
- The model fits the data (By Hosmer and Lemeshow test)
- The factors age, gender, education, food habits and topography sites are non-significant.
- The rate of lives is 0.359 times lower in null model. In other words, the mortality rate is higher in null model (i.e. the factors are not assigned)
- The mortality rate increases with the increase of behavioural factors (Consumption of areca nut, tobacco, smoking and alcohol)
- The mortality rate is lower in higher family income groups.
- The mortality rate is more when CDT (Cancer directed treatment) is not assigned or assigned but not detected.
- The mortality rate is quite more if the treatment is not given prior to registration.
- the mortality rate increases in higher metabolic risk factors (over weight, raised blood pressure, raised blood glucose)
- The mortality rate is higher in urban areas.
- There are eight outliers by using the model. Among them there seven are false predictions and one is true prediction.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram M, Jemal A, *et al*. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021;71(3):209-49. https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21660
2. Assam-National Centre for Disease Informatics and Research. Cancer in North-East Region of India: A Report of Population Based Cancer Registries 2012-2016. Bengaluru: NCDIR-ICMR; 2021. https://ncdirindia.org/All_Reports/NorthEast2021/resources/NE_chapter3.pdf
3. Bagley SC, White H, Golemb BA. Logistic regression in the medical literature: Standards for use and reporting with particular attention to one medical domain. Journal

of Clinical Epidemiology. 2001;54(10):979-985. [suspicious link removed]

4. Koç Z, Çelebi P, Memiş A, Sağlam Z, Beyhan F. Evaluation of Impact of Nurses' Healthy Lifestyle Behaviours on Utilization from Breast Cancer Early Diagnosis Methods. Journal of Breast Health. 2014;10(1):1-7. https://ncbi.nlm.nih.gov/pmc/articles/pmc5351542
5. Mahana B Madana B. A study to compare critically the prognosis of malignant cancer patients through various statistical models [dissertation]. [Coimbatore]: Bharathiar University; 2008. http://hdl.handle.net/10603/63012
6. Al-Hakimi HA, Othman AE, Al-Khateeb AA. Public Knowledge of Oral Cancer and Modelling of Demographic Background Factors Affecting this Knowledge in Khartoum State, Sudan. International Journal of Preventive Medicine. 2016;7:118. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4996297/
7. Hung LC, Kung PT, Pan LY, Lin YJ, Tsai SY, Lee CL, *et al*. Assessment of the Risk of Oral Cancer Incidence in a High-Risk Population and Establishment of a Prediction Model for Oral Cancer Incidence Using a Population-Based Cohort in Taiwan. International Journal of Environmental Research and Public Health. 2020;17(2):644. https://pubmed.ncbi.nlm.nih.gov/31968579/
8. Schwartz SM, Daling JR. Oral Cancer Risk in Relation to Sexual History and Evidence of *Human Papillomavirus* Infection. Journal of the National Cancer Institute. 1998;90(21):1626-1636. https://academic.oup.com/jnci/article/90/21/1626/2520309
9. Prasad SN, Diwakar AK. Statistical Analysis of Cancer Data. International Journal of Scientific Research. 2022;11(8):12-13. https://ijsr.net/archive/V11i8/SR22819190317.pdf
10. Bharathi A, Natarajan AM. Cancer Classification of Bioinformatics Data using ANOVA. International Journal of Computer Theory and Engineering. 2223;2(5):789-793. https://ijcte.org/papers/169-G652.pdf
11. Rana R, Singhal R. Chi-square Test and its Application in Hypothesis Testing. Journal of the Practice of Cardiovascular Sciences. 2015;1(1):69-72. https://j-pcs.org/temp/JPractCardiovascSci1169-5772018_160200.pdf
12. Adimu PI, Oguntunde PE, Okagbue HI, Agboola O. Statistical data analysis of cancer incidences in insurgency affected states in Nigeria. Heliyon. 2018;4(6):e00667. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5998707/
13. Zangmo C, Tiensuwan M. Application of logistic regression models to cancer patients: a case study of data from Jigme Dorji Wangchuck National Referral Hospitals (JDWNRH) in Bhutan. Journal of Physics: Conference Series. 2018;1039(1):012031. https://iopscience.iop.org/article/10.1088/1742-6596/1039/1/012031/pdf
14. Adedin T, Chowdhury ZI, Afzal A, Yeasmin F. Application of Binary Logistic Regression in Clinical Research. Bangladesh Journal of Medical Science. 2016;15(4):596-600. https://www.researchgate.net/publication/320432727_Application_of_Binary_Logistic_Regression_in_Clinical_Research
15. Suresh N, Thomas B, Joseph J. Bibliometric Analysis and Visualization of Scientific Literature on Heart Disease Classification using Logistic Regression Model. Cureus. 2024;16(7):e64765. https://assets.cureus.com/uploads/review_article/pdf/264765/20240728-1125897-xxq9jt.pdf
16. Aziz H. Assessing the Performance of Logistic Regression in Heart Disease Detection through 5-Fold Cross-Validations. International Journal of Artificial Intelligence in Medical Issues. 2024;2(1):31-36. https://jurnal.yoctobrain.org/index.php/ijaimi/article/view/137/187
17. Alwiah I, Umar Z, Murdiyanto. Assessing the predictive Power of Logistic Regression on Liver Disease Prevalence in Indian Context. Indonesian Journal of Data and Science. 2024;5(1):1-8. https://jurnal.yoctobrain.org/index.php/ijodas/article/view/121/172
18. Amin MB, Edge SB, Greene FL, Byrd DR, Brookland RK, Washington MK, *et al*., editors. AJCC Cancer Staging Manual. 8th ed. New York: Springer; 2017.
19. Shaker AA, Fahem NA, Hayder M. Predicting Diabetes Disease Occurrence Using Logistic Regression: An Early Detection Approach. Iraqi Journal for Computer Science and Mathematics. 2024;5(2):167-176. https://www.iraqoaj.net/iasj/download/e54cced8433b3098