Version No.: 4.0 **Date: 2026.1.26**

Machine Learning-Based Risk Prediction for Head and Neck Cancerous Lesions: A Multidimensional Feature Study Based on Demographics and Clinical Symptomatology

Principal Investigator: Chuanyao Lin

Department: Department of Otorhinolaryngology-Head and Neck Surgery

Contact Person: Junyi Wang

Contact Phone: 13309609232

Study Period: April 2026 to November 2030

Version: 4.0

Version Date: January 26, 2026

Nanjing Drum Tower Hospital, Affiliated Hospital of Nanjing University Medical School

Project Title	Machine Learning-Based Risk Prediction for Head and Neck Cancerous Lesions: A Multidimensional Feature Study Based on Demographics and Clinical Symptomatology
Study Objective	This study aims to utilize machine learning algorithms based on multi-center clinical data to develop and validate a clinical prediction model for assessing the risk of head and neck cancerous lesions.
Study Design	Study Process: (1) A retrospective cohort study design is utilized for the preliminary construction and internal performance evaluation of the model. (2) A prospective observational cohort study design is employed to evaluate the generalization ability and clinical applicability of the final model.
Total Number of Cases	11,000 (Retrospective study: 8,000 cases; Prospective study: 3,000 cases)
Patient Selection	Inclusion Criteria: Patients presenting at the Department of Otorhinolaryngology-Head and Neck Surgery, Nanjing Drum Tower Hospital, who undergo laryngoscopy, sign the informed consent form, and possess complete clinical data.
	Exclusion Criteria: Patients with incomplete clinical data.
Treatment Regimen	None

Outcome Assessment	Efficacy Evaluation Indicators (Primary and Secondary Efficacy Indicators): As this is an observational predictive modeling study that does not involve interventions, there are no traditional efficacy or safety indicators.
	Safety Evaluation Indicators: As this is an observational predictive modeling study that does not involve interventions, there are no traditional efficacy or safety indicators.
Statistical Methods	The Mann-Whitney U test is utilized to analyze the test results between the non-diseased group and the diseased group.
Study Period	April 2026 to November 2030

I. Study Background

Head and neck malignancies (primarily cancerous lesions of the oral cavity, pharynx, and larynx) are the sixth most common cancer globally. In recent years, their incidence has continued to climb, posing a severe public health challenge [1]. Due to their deep anatomical locations, early clinical symptoms (such as persistent hoarseness, foreign body sensation in the throat, and non-healing oral ulcers) are insidious and lack specificity. Clinically, these symptoms are highly similar to those of benign conditions such as chronic inflammation, easily leading to missed diagnoses, misdiagnoses, and delayed treatment [2]. Epidemiological surveys show that over 60% of patients have already progressed to a locally advanced stage or present with cervical lymph node metastasis at the time of initial diagnosis. Advanced-stage cancer patients not only face complex treatment strategies, significant surgical trauma, and heavy financial burdens, but they also suffer severe sequelae, such as permanent impairment of speech and swallowing functions, leading to a precipitous drop in quality of life and a significantly reduced five-year survival rate [3].

Currently, the clinical diagnostic pathway for head and neck cancers relies heavily on a series of invasive and resource-intensive examinations. Electronic laryngoscopy can visually assess mucosal lesions, but the procedure easily triggers the gag reflex and other discomforts, resulting in poor patient tolerance, and it relies heavily on the specialist's operating experience [4]. Imaging technologies such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) are vital adjunctive diagnostic tools [5] that provide valuable anatomical information, but they involve high equipment costs and expensive examination fees. The accessibility of such examinations as primary screening tools for head and neck malignancies in grassroots medical institutions is limited [6]. Furthermore, histopathological biopsy, the gold standard for diagnosis, is an invasive procedure that is not only technically demanding but also causes physical and psychological distress to patients, carrying potential risks of complications [7]. These combined factors constrain the feasibility of

implementing large-scale, accessible early screening in communities and high-risk populations, leading to an inefficient allocation of medical resources and a substantial proportion of missed and delayed diagnoses in early-stage cases.

Given this background, clinical practice urgently requires the development of a simple, reliable, and clinically applicable primary screening risk stratification tool based on routine clinical variables, demographic data, and symptom questionnaires through multidimensional analysis. This tool would enable simple preliminary screening of high-risk populations for head and neck cancer, thereby optimizing prevention and treatment strategies and improving patient prognosis [8]. In recent years, the rapid development of machine learning technology has provided a new methodological approach to achieving this goal. Compared to traditional statistical models, machine learning can handle multidimensional, non-linear clinical data and automatically identify complex interactions and hidden patterns among variables, thereby building predictive models with superior discriminatory performance [6]. Therefore, this study plans to systematically integrate easily accessible clinical data, including demographic characteristics, lifestyle indicators, and clinical manifestations. Utilizing machine learning algorithms such as Random Forest and XGBoost, we aim to develop and validate a highly accurate and generalizable risk prediction model for head and neck cancerous lesions. This model will facilitate preliminary screening in the early stages of the disease, guide high-risk individuals to undergo timely specialist examinations, and promote multidisciplinary joint diagnosis and treatment, effectively improving patients' quality of life and preventing complications [9].

II. Study Objectives

1. Primary Objective This study aims to utilize machine learning algorithms based on multi-center clinical data to develop a low-cost, non-invasive primary screening risk stratification tool to identify individuals at high risk for head and neck malignancies and optimize early diagnostic and treatment pathways.

2. Secondary Objectives

Identify and quantify the independent and interactive effects of key demographic information, lifestyle and clinical characteristics, and laryngoscopy reports on the risk of developing head and neck cancerous lesions.

Systematically compare the predictive performance of various mainstream machine learning algorithms (e.g., Logistic Regression, Random Forest, Gradient Boosting Machine) within this study cohort to select the optimal modeling strategy.

Apply model interpretability techniques, such as SHAP (SHapley Additive exPlanations), to analyze the final primary screening risk stratification tool, elucidate its decision-making rationale, and enhance the clinical transparency and credibility of the model.

Evaluate the clinical applicability of the model using Decision Curve Analysis (DCA) to verify its clinical efficacy in assisting early screening and managing high-risk populations.

Evaluate the reliability and validity of a composite multidimensional prediction model incorporating laryngoscopy results compared to purely symptomatological and demographic models, to optimize screening protocols for patients with head and neck cancerous lesions.

III. Study Design Types, Principles, and Trial Procedures

1. Study Design This study adopts a multi-center, observational study design, adhering to the TRIPOD reporting guidelines, and is divided into two sequential phases.

Phase One (Model Development & Internal Validation):

Study Design Type: Retrospective cohort study

Data Source: Nanjing Drum Tower Hospital, Shanghai Ninth People's Hospital, Changzhou No. 4 People's Hospital, and Cixi People's Hospital (2018–2025).

Sample Size: 8,000 cases. The data will be split at a 7:3 ratio into a training set (n=5,600) and an internal validation set (n=2,400) for preliminary model construction and internal performance evaluation.

Phase Two (External Model Validation):

Study Design Type: Prospective observational cohort study

Data Source: Nanjing Drum Tower Hospital, Shanghai Ninth People's Hospital, Changzhou No. 4 People's Hospital, and Cixi People's Hospital et al.(2026–2030).

Sample Size: 3,000 cases (competitive enrollment across centers).

Objective: To evaluate the generalizability and temporal stability of the model in new populations.

2. Study Centers and Sample Size This study will be conducted across four centers: Nanjing Drum Tower Hospital, Shanghai Ninth People's Hospital, Changzhou No. 4 People's Hospital, and Cixi People's Hospital. The retrospective development cohort plans to randomly enroll 8,000 outpatient cases meeting the inclusion/exclusion criteria (competitive enrollment per center). The prospective validation cohort plans to consecutively enroll 3,000 outpatient patients meeting the inclusion/exclusion criteria (competitive enrollment per center).

3. Study Procedures This study will be systematically carried out in the following steps:

Step 1: Data Extraction and Preprocessing

Data Collection: Retrospective data will be extracted from the Electronic Medical Record (EMR) systems of the four participating centers strictly according to the inclusion and exclusion criteria. Data collected includes:

Demographics: Age (years), Gender, Education level (middle school or below / high school / college or above), Occupation (manual labor / non-manual labor / retired).

Lifestyle: Smoking index (pack-years = packs per day × years smoked), Alcohol consumption (yes/no; frequency/amount), History of betel nut chewing, Preference for spicy/hot food.

Medical History: HPV infection status (serum HPV16/18 antibodies or oropharyngeal swab PCR results), Reflux gastritis (gastroscopy or clinical diagnosis), Diabetes, Immunosuppressive status.

Symptom Characteristics: Type of chief complaint (hoarseness, odynophagia, foreign body sensation, neck mass, etc.), Duration (< 2 weeks / ≥ 2 weeks), Presence of progressive worsening.

Laryngeal Questionnaires: Reflux Symptom Index (RSI), Voice Handicap Index-10 (VHI-10).

Electronic Laryngoscopy: Laryngoscopy result report (normal / inflammation / suspicious neoplasm / definite space-occupying lesion; unilateral/bilateral; vocal cord involvement).

Data Cleaning & Preprocessing: Clean the collected data and handle missing values using multiple imputation methods. Apply one-hot encoding for categorical variables and standardize continuous variables in preparation for machine learning modeling.

Data Grouping: Based on laryngoscopy reports and subsequent surgical biopsy pathology reports, patients will be divided into a diseased group and a non-diseased group for downstream data processing and analysis.

Step 2: Feature Engineering and Dataset Splitting

Feature Selection: On the retrospective training set, feature selection methods such as LASSO regression will be utilized to identify predictor variables most relevant to the outcome, optimizing the model and preventing overfitting.

Dataset Splitting: The complete retrospective cohort data will be randomly divided into a training set and an internal testing set at a 7:3 ratio.

Step 3: Machine Learning Model Development and Internal Validation

Model Training: Simultaneously train multiple machine learning algorithms on the training set, including but not limited to Logistic Regression, Random Forest, XGBoost, and Support Vector Machines (SVM).

Hyperparameter Tuning: Employ five-fold cross-validation and grid search techniques to find the optimal hyperparameter combinations for each algorithm.

Internal Performance Evaluation: Evaluate the performance of each candidate model on the locked internal testing set. Primary evaluation metrics include the Area Under the ROC Curve (AUC), Accuracy, Recall, F1-score, and Calibration curves.

Step 4: Model Interpretation and Finalization

Model Interpretation: For the best-performing model, utilize model interpretability techniques like SHAP to quantify the contribution of each feature to individual predictions, generating global and individual-level explanation plots.

Model Finalization: Comprehensively consider the model's discrimination, calibration, and clinical interpretability to finalize the prediction model.

Step 5: Prospective External Validation and Clinical Utility Evaluation

Prospective Validation: Prospectively and consecutively collect new case data from two (or more) centers within a predefined timeframe to form an independent external validation cohort.

Performance Reproduction Evaluation: Validate the final model on this cohort to evaluate its generalization performance.

Clinical Utility Evaluation: Employ Decision Curve Analysis (DCA) to assess the clinical net benefit of the model across different decision thresholds, clarifying its practical value in assisting clinical decision-making.

IV. Case Selection

1. Sample Size This study will be conducted across four centers: Nanjing Drum Tower Hospital, Shanghai Ninth People's Hospital, Changzhou No. 4 People's Hospital, and Cixi People's Hospital. The retrospective development cohort plans to randomly enroll 8,000 outpatient cases meeting the inclusion and exclusion criteria (competitive enrollment per center). The prospective validation cohort plans to consecutively enroll 3,000 outpatient patients meeting the inclusion and exclusion criteria (competitive enrollment per center).

2. Study Groups:

Positive (Diseased Group): Laryngoscopy finding of a suspicious lesion + subsequent pathological confirmation of Head and Neck Squamous Cell Carcinoma (HNSCC).

Negative (Non-Diseased Group): Negative laryngoscopy results or solely benign lesions (e.g., vocal cord polyps, pharyngeal/laryngeal cysts) with no confirmed malignancy over a follow-up period of at least 6 months.

3. Definition of Head and Neck Malignancy:

Anatomical Scope: Oral cavity (tongue body, gingiva, buccal mucosa, floor of the mouth, hard palate), Oropharynx (tonsils, soft palate, base of tongue), Hypopharynx, Larynx (supraglottis, glottis, subglottis).

Histological Type: Primary Squamous Cell Carcinoma (including subtypes like verrucous carcinoma, basaloid squamous cell carcinoma).

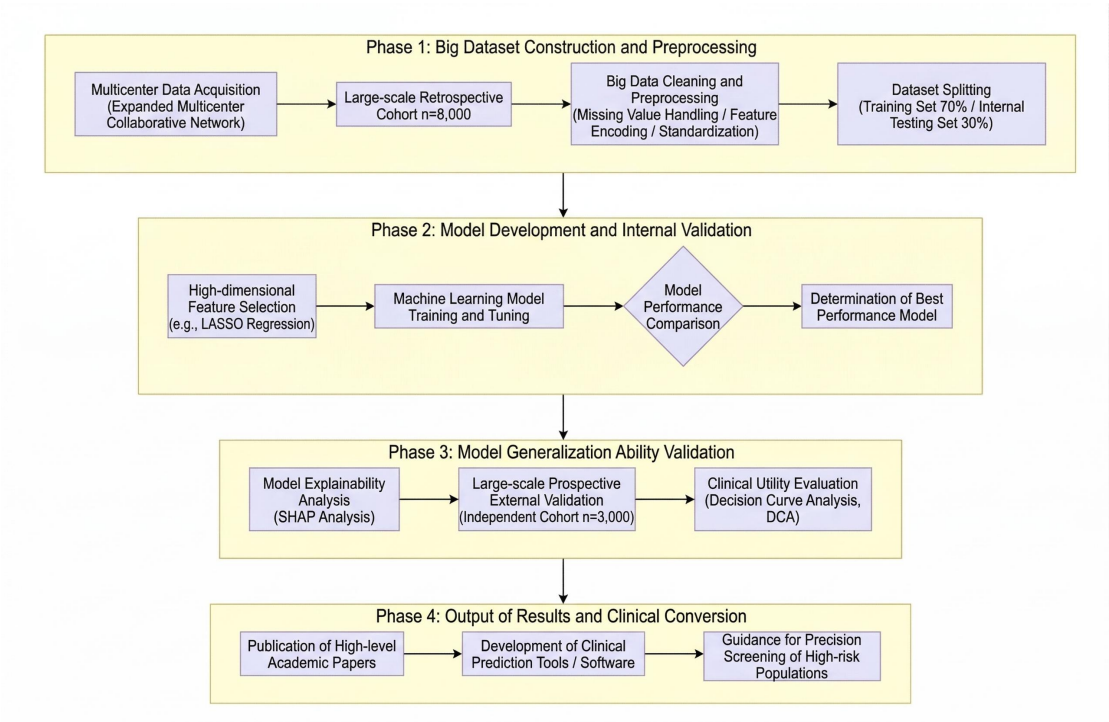
Exclusions: Nasopharyngeal carcinoma (different EBV-related mechanism), Thyroid cancer, Salivary gland adenocarcinoma, Lymphoma, Metastatic cancer, Skin cancer, Precancerous lesions.

Gold Standard: Confirmed via pathological biopsy or postoperative specimen, with reports issued by the pathology departments of the collaborating hospitals.

4. Inclusion Criteria Patients enrolled at the study centers must be those presenting at the Department of Otorhinolaryngology-Head and Neck Surgery at Nanjing Drum Tower Hospital (and collaborating centers) who have undergone a laryngoscopy. For the retrospective study, informed consent is waived; for the prospective study, patients must sign an informed consent form. Patients in both the retrospective and prospective arms must have complete clinical data.

5. Exclusion Criteria Patients who refuse to sign the informed consent form (for the prospective arm), have incomplete clinical data, or have known other head and neck malignancies (e.g., thyroid cancer, malignant parotid gland tumors, etc.).

V. Flow-Chart



VI. Study Quality Control and Quality Assurance

The Department of Otorhinolaryngology-Head and Neck Surgery at Nanjing Drum Tower Hospital has established a specialized quality control group. Procedures ranging from data sample collection and testing to data analysis will be strictly implemented in accordance with relevant Standard Operating Procedures (SOPs). The Principal Investigator and all relevant members have received professional skills training.

VII. Data Safety Monitoring

A corresponding data safety monitoring plan will be formulated for the clinical study based on the level of risk. Any personal information and data obtained about you during the trial will be kept strictly confidential. Your information, data, and questionnaire information will be identified by a study number/digit rather than your name. Information that can identify your identity will not be disclosed to members outside the research team without your permission. All research members and the

sponsor are required to maintain confidentiality regarding your identity. Your files will be stored in a locked data cabinet, accessible only to research personnel. To ensure the study is conducted according to regulations, when necessary, government regulatory authorities, monitors authorized by the sponsor, or Ethics Committee members may, according to regulations, access relevant information concerning your participation in the trial at the research site; however, they guarantee not to disclose your information to other parties. Although study results may be published, your identity will not be revealed in these publications. All detailed data of this study will be stored at Nanjing Drum Tower Hospital, Affiliated Hospital of Nanjing University Medical School.

VIII. Ethics of Clinical Research

The clinical study will adhere to relevant regulations, including the World Medical Association's "Declaration of Helsinki." Clinical research shall only be implemented after the Ethics Committee has approved the trial protocol before the start of the study. Before enrolling each subject into this study, the investigator is responsible for providing a complete and comprehensive overview of the purpose, procedures, and potential risks of this study to the subject or their legal representative, and for signing a written informed consent form. Subjects should be informed of their right to withdraw from the study at any time. The informed consent form should be retained as a clinical research document for review. Subject personal privacy and data confidentiality will be protected throughout the research process.

IX. Collection and Storage of Samples/Specimens

All clinical data will be managed by specialized personnel to ensure data privacy, confidentiality, and security.

X. Study Timeline

April 2026 – November 2029: Prospectively collect data from 3,000 patients undergoing laryngoscopy examinations at the Otorhinolaryngology-Head and Neck Surgery outpatient clinics of Nanjing Drum Tower Hospital, Changzhou No. 4 People's Hospital, Cixi People's Hospital, and Shanghai Ninth People's Hospital.

November 2028 – November 2029: Complete all sample collection and data analysis.

November 2029 – November 2030: Machine learning modeling, patent application, and manuscript writing.

XII Reference

1. FANG E F, XIE C, SCHENKEL J A, et al. A research agenda for ageing in China in the 21st century (2nd edition): focusing on basic and translational research, long-term care, policy and social networks[J]. Ageing Res Rev, 2020, 64: 101174. DOI: 10.1016/j.arr.2020.101174.

2. ABDI S, SPANN A, BORILOVIC J, et al. Understanding the care and support needs of older people: a scoping review and categorisation using the WHO international classification of functioning, disability and health framework (ICF)[J]. BMC Geriatr, 2019, 19(1): 195. DOI: 10.1186/s12877-019-1189-9.

3. LECHIEN J R, GENEID A, BOHLENDER J E, et al. Consensus for voice quality assessment in clinical practice: guidelines of the European Laryngological Society and Union of the European Phoniatricians[J]. Eur Arch Otorhinolaryngol, 2023, 280(12): 5459–5473. DOI: 10.1007/s00405-023-08211-6.

4. KIM T, CHOI J Y, KO M J, KIM K I. Development and validation of a machine learning method using vocal biomarkers for identifying frailty in community-dwelling older adults: cross-sectional study[J]. JMIR Med Inform, 2025, 13: e57298. DOI: 10.2196/57298.

5. LIN Y C, YAN H T, LIN C H, CHANG H H. Identifying and estimating frailty phenotypes by vocal biomarkers: cross-sectional study[J]. J Med Internet Res, 2024, 26: e58466. DOI: 10.2196/58466.

6. Bray F ,Laversanne M ,Sung H ,et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA Cancer J Clin, 2024,74(3):229-263. DOI: 10.3322/caac.21834 .

7. HOU Y J, YANG X X, MENG H X. Mitochondrial metabolism in laryngeal cancer: therapeutic mechanisms and prospects[J]. Biochim Biophys Acta Rev Cancer, 2025, 1880(3): 189335. DOI: 10.1016/j.bbcan.2025.189335.

8. Govil N ,Tripathi M ,Parag K ,et al. Role of protocol-guided perioperative care to enhance recovery after head and neck neoplasm surgery: an institutional experience[J]. Rev Esp Anestesiol Reanim (Engl Ed), 2023,70(9):491-500. DOI: 10.1016/j.redare.2023.09.002 .