

A Simplified Nomogram for Primary Population-Based Screening of Esophageal Cancer: An Internally Validated Shandong Cohort Study Addressing Data Imbalance

Junjun Hou^{1,*}, Xiuyu He^{2,*}, Fei Gao², Ting Ren², Lijin Ren³, Xianguo Wang⁴

¹Department of Medical Oncology I, Tai'an Cancer Hospital, Tai'an, Shandong, 271099, People's Republic of China; ²Department of Office of Cancer Center, Tai'an Cancer Hospital, Tai'an, Shandong, 271099, People's Republic of China; ³Department of Traditional Chinese Medicine I, Tai'an Cancer Hospital, Tai'an, Shandong, 271099, People's Republic of China; ⁴Department of Medical Oncology II, Tai'an Cancer Hospital, Tai'an, Shandong, 271099, People's Republic of China

*These authors contributed equally to this work

Correspondence: Xianguo Wang, Department of Medical Oncology II, Tai'an Cancer Hospital, No. 390, Lingshan Street, Taishan District, Tai'an, Shandong, 271099, People's Republic of China, Tel +86 13562898761, Email Wangxianguo55@163.com; Lijin Ren, Department of Traditional Chinese Medicine I, Tai'an Cancer Hospital, No. 390, Lingshan Street, Taishan District, Tai'an, Shandong, 271099, People's Republic of China, Email renlijing3@21cn.com

Objective: To develop and validate a user-friendly nomogram for predicting the risk of esophageal squamous cell carcinoma (ESCC) and high-grade intraepithelial neoplasia (HGIN), designed for initial screening settings while addressing variable complexity and class imbalance in traditional models.

Methods: Based on a screening cohort of 23,257 participants from Tai'an, Shandong, between 2024 and 2025 (positive rate: 1.54%), a 1:10 case-control sampling method was applied to address the low event rate (positive rate: 1.54%) and correct class imbalance. Predictors were initially screened using LASSO regression with 10-fold cross-validation ($\lambda_{\min}$ criterion) and further refined via multivariable logistic regression to establish the final model, which was presented as a nomogram and evaluated via internal split-sample validation.

Results: Seven easily accessible predictors were identified: age, sex, education level, BMI, smoking history, hot-food consumption, and family history of esophageal cancer. The model showed strong discriminatory performance, with an AUC of 0.823 (95% CI: 0.798–0.848) in the training set and 0.835 (95% CI: 0.805–0.865) in the internal validation set. Calibration curves indicated high consistency between predicted and observed risks. Decision curve analysis demonstrated net clinical benefit across risk thresholds of 0–0.6.

Conclusion: The proposed simplified nomogram demonstrates promising potential for risk stratification in primary ESCC/HGIN screening. However, prospective external validation in diverse cohorts is necessary before its large-scale clinical implementation.

Keywords: esophageal squamous cell carcinoma, precancerous conditions, early detection of cancer, predictive learning models

Introduction

According to the latest GLOBOCAN 2022 statistics, esophageal cancer remains one of the most prevalent malignancies worldwide, with China bearing a disproportionately high share of both incidence and mortality. Recent epidemiological data indicate that nearly half of newly diagnosed cases and deaths globally occur in China, with rural regions experiencing a particularly heavy disease burden.^{1,2} Esophageal squamous cell carcinoma (ESCC), a predominant histological subtype in China, is characterized by an unfavorable prognosis and insidious early manifestations.² Given the absence of specific symptoms in the early stage, most patients are diagnosed at advanced or metastatic stages, when the 5-year survival rate is approximately 30.3%.³ In contrast, patients treated at an early stage can achieve 5-year survival rates exceeding 80%.⁴ These disparities highlight the importance of early detection and timely intervention in reducing ESCC-related mortality, with screening representing a cornerstone of prevention strategies.^{5,6}

Endoscopy is widely regarded as the reference standard for esophageal cancer screening. However, its application in large-scale population screening is constrained by high costs, invasiveness, and suboptimal adherence. Furthermore, the distribution of medical resources in China remains markedly uneven. High-quality endoscopic equipment and experienced endoscopists are largely available only in tertiary hospitals in urban centers, whereas primary healthcare institutions often lack the necessary infrastructure and technical capacity. These limitations hinder the implementation of universal endoscopic screening. Accordingly, developing risk prediction models based on readily obtainable and routinely collected risk factors may enable risk stratification within the general population, thereby reducing unnecessary endoscopic procedures and improving screening efficiency.⁷

In recent years, several ESCC risk prediction models have been proposed using hospital-based or community-based cohorts, including recent models based on endoscopic characteristics⁸ or electronic health records.⁹ Nevertheless, notable shortcomings persist, leaving a significant gap in practical primary screening tools. First, some models incorporate complex or difficult-to-access variables, including genetic testing results or endoscopic pathological features,⁸ which restrict their feasibility for large-scale screening prior to endoscopy. Second, most prior studies have not adequately addressed the issue of outcome imbalance inherent in screening datasets,⁷ where positive cases are exceptionally rare. This severe imbalance is a central methodological issue because it potentially compromises predictive performance and introduces bias toward the majority class (screen-negative individuals). Therefore, the true novelty of our study lies in specifically addressing this gap by developing a simplified nomogram based entirely on easily accessible, non-invasive questionnaire data, while integrating a rigorous class-imbalance correction method to ensure accurate risk ranking for rare outcomes.

Using data derived from an upper gastrointestinal cancer screening program conducted in Tai'an, Shandong Province, this study developed and internally validated a risk prediction model for ESCC and high-grade intraepithelial neoplasia (HGIN), aiming to refine primary screening strategies and offer methodological insight for handling class imbalance in cancer risk prediction studies. To mitigate bias arising from rare-event imbalance, we implemented a 1:10 case-control sampling scheme to recalibrate the dataset before model development.

Study Population

The overall study design and the participant selection process are summarized in the flowchart ([Supplementary Figure 1](#)). This retrospective analysis was conducted within a population-based screening cohort derived from the “Opportunistic Screening Program for Early Detection and Early Treatment of Upper Gastrointestinal Cancer” in Tai'an, Shandong Province, between January 2024 and January 2025. This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Tai'an Cancer Hospital (Tai'an Cancer Center, Tai'an Third People's Hospital) (Approval No: 232016). Written informed consent was obtained from all participants before enrollment.

Inclusion criteria were as follows: (1) age 40–69 years; (2) no prior history of malignant tumors; and (3) completion of both the structured questionnaire and endoscopic examination. A total of 26,439 individuals were initially included.

Exclusion criteria were as follows: (1) duplicate records; (2) missing data $\geq 30\%$ in key variables; and (3) a baseline diagnosis of HGIN. Key variables were defined a priori and included demographic characteristics (*ie.*, sex, age, educational level), core clinical and lifestyle factors (*ie.*, body mass index [BMI], smoking history, preference for very hot beverages or foods, family history of cancer), and outcome measures (*ie.*, endoscopic findings, histopathological diagnosis).

After data quality control procedures, 2,677 participants were excluded due to duplicate entries or excessive missing data in key variables, and an additional 505 individuals with HGIN at baseline were removed. Ultimately, 23,257 participants were retained for analysis. Among them, 358 cases were histopathologically confirmed as ESCC or HGIN, corresponding to a detection rate of 1.54%.

Given the low detection rate of esophageal cancer and precancerous lesions in the original dataset (1.54%), direct model development using the full cohort would likely bias the classifier toward the majority class (screen-negative individuals). To mitigate this class imbalance, we adopted a case-control design following the methodological framework proposed by Fu et al¹⁰ for handling rare outcomes. Compared to synthetic data generation techniques (*eg.*, SMOTE) which may introduce artificial data noise, this resampling strategy relies entirely on real-world clinical data, ensuring greater clinical interpretability. Specifically, a 1:10 case-control sampling strategy was implemented to rebalance the dataset. The selection of this ratio was guided by several considerations. First, previous methodological studies, including that of Fu et al,¹⁰ have

suggested that when the outcome prevalence is extremely low (<5%), a 1:10 ratio can enhance the model's capacity to capture characteristics of the minority class more effectively than conventional ratios such as 1:1 or 1:4. Second, findings from preliminary analyses indicated that, compared with a 1:5 ratio, the 1:10 design yielded more stable regression coefficient estimates, as reflected by smaller standard errors. Finally, relative to higher ratios (eg, 1:20), the 1:10 approach preserved statistical power while avoiding unnecessary redundancy among controls, thereby reducing computational burden and the risk of overfitting. All 358 participants with histopathologically confirmed ESCC or HGIN were included in the case group. Controls were selected using simple random sampling. Specifically, a computer-generated randomization procedure implemented in R was applied to the 22,899 screen-negative individuals, from whom 3,580 were randomly drawn as controls. This computer-assisted random sampling ensured that the control group was selected in an unbiased and representative manner. Importantly, while this design optimizes risk ranking (discrimination) for stratified screening, it alters the baseline prevalence. Therefore, the resulting predicted probabilities should be interpreted as relative risk scores used for clinical risk stratification, rather than true absolute risks in the general population.

Assessment of Risk Factors

Baseline information for all eligible participants was extracted from the information management system of the "Opportunistic Screening Program for Early Detection and Early Treatment of Upper Gastrointestinal Cancer" in Tai'an, Shandong Province. In the original screening program, data were collected through face-to-face interviews conducted by trained investigators. A series of standardized procedures were implemented to ensure data reliability. First, a unified questionnaire was used, developed by the Expert Committee of the Upper Gastrointestinal Cancer Panel under the Rural Cancer Early Diagnosis and Early Treatment Program of the China Cancer Foundation. Second, before study initiation, all interviewers underwent structured training covering protocol interpretation, operational definitions of each questionnaire item, and interview techniques. During data collection, standardized instructions were used to minimize interviewer-induced variability. Third, dedicated quality control personnel were assigned at each screening site to perform daily logical checks and verification of completed questionnaires. Variables retrieved for this analysis included demographic characteristics (*ie.*, age, sex, educational level), anthropometric measures (*ie.*, height and weight, used to calculate BMI), lifestyle factors (*ie.*, smoking history, alcohol consumption, tea drinking habits), dietary habits (*ie.*, frequency of consumption of very hot food or beverages and seven additional dietary frequency categories), and history of gastrointestinal diseases and family history of esophageal cancer.

Statistical Analysis

All statistical analyses were performed using the software R 4.5.0. The analytical workflow comprised the following components:

- (1) Handling of missing data: Missing values in continuous variables, such as height and weight, were addressed using multiple imputations with predictive mean matching, generating five complete datasets. The imputation model incorporated demographic variables, all lifestyle and dietary variables, and pathological outcomes to preserve the underlying correlation structure among variables. Estimates from the imputed datasets were combined according to Rubin's rules.
- (2) Descriptive analysis: Categorical variables were summarized as frequencies and percentages ($n[\%]$). Continuous variables (*ie.*, age, BMI) were assessed for normality using Q-Q plots and histograms and were presented as mean $\pm$ standard deviation ($\bar{x} \pm s$).

Assessment of associations: Restricted cubic spline (RCS) regression was applied to explore potential nonlinear associations between continuous variables (*ie.*, age, BMI) and the outcome. Four knots were used, located at the 5th, 35th, 65th, and 95th percentiles.

- (3) Outcome definition: The primary endpoint was the presence of HGIN or more advanced lesions (HGIN+), confirmed by histopathological examination of endoscopic biopsy specimens or surgically resected tissue.

Diagnostic criteria were defined in accordance with the World Health Organization Classification of Tumours of the Digestive System. HGIN was characterized by marked cytologic and architectural atypia involving the full thickness of the epithelium, with preservation of the basement membrane. Carcinoma in situ was defined as full-thickness epithelial involvement by atypical cells exhibiting unequivocal malignant cytological features. ESCC was diagnosed when atypical squamous cells breached the basement membrane and invaded the lamina propria, submucosa, or deeper layers.

- (4) Sample size estimation: Sample size adequacy was evaluated based on the conventional Events Per Variable (EPV) principle for prediction modeling. To ensure stable coefficient estimation in logistic regression and to minimize overfitting, an EPV of at least 10 is generally recommended. In the present study, 358 events were observed in the case group, and 25 candidate predictors were initially considered. This yielded an EPV of 14.32 (358/25), exceeding the recommended threshold. These calculations indicate that the available sample size was sufficient to support the reliable development of a multivariate risk prediction model.
- (5) Dataset partitioning: Using the built-in random number generator in R (with a predefined random seed to ensure reproducibility), the resampled dataset was randomly divided into a training set and an internal validation set at a ratio of 8:2. Stratified random sampling was applied to maintain comparable proportions of outcome events between the two subsets.
- (6) Predictor selection: Variable selection was performed within the training set. Least absolute shrinkage and selection operator (LASSO) regression with 10-fold cross-validation was first applied for preliminary screening. The optimal penalty parameter (λ) was determined according to the $\lambda_{\min}$ criterion, corresponding to the minimum cross-validated prediction error. Subsequently, multicollinearity diagnostics were conducted by calculating variance inflation factors (VIFs), with a predefined threshold of $VIF < 5$. Predictors meeting this criterion were entered into a multivariate logistic regression model. Variables with two-sided P-values < 0.05 were retained in the final model, whereas those with P-values exceeding 0.10 were excluded.
- (7) Model development and validation: A multivariate logistic regression model was constructed based on the final set of predictors, and a nomogram was generated to facilitate individualized risk estimation. Model performance was evaluated in both the training and validation sets across three dimensions, namely, discrimination, calibration, and clinical utility. Discrimination was assessed using the area under the receiver operating characteristic curve (AUC). Calibration was examined through calibration plots with 1,000 bootstrap resamples. Clinical utility was evaluated by decision curve analysis (DCA).
- (8) Clinical translation: For practical implementation, the optimal risk threshold was determined using Youden's index. Participants were subsequently categorized into high- and low-risk groups according to this cutoff. The potential reduction in healthcare resource utilization was estimated by calculating the proportion of endoscopic examinations that could be avoided compared with a universal endoscopic screening strategy.

Results

Baseline Characteristics

Nonlinearity was assessed using the Wald test. The results indicated no significant deviation from linearity for either age or BMI (both $P > 0.05$, [Figure 1](#)), supporting their inclusion as continuous variables in subsequent regression analyses.

A total of 3,938 participants were included in the final analytic sample. Based on endoscopic histopathological findings, subjects were classified into two groups: 3,580 individuals with lesions below HGIN and 358 individuals with HGIN or above. Comparative analyses between the two groups revealed statistically significant differences in age, BMI, sex, educational level, smoking history, alcohol consumption, intake of fresh vegetables, intake of fresh fruit, and consumption of pickled foods, fried foods, very hot foods or beverages, and mold-contaminated foods (all $P < 0.05$). See [Table 1](#). For example, alcohol consumption differed significantly between groups ($P = 0.003$), with a lower proportion of individuals reporting a history of alcohol use in the HGIN or above group (71.79%) compared with the control group (78.85%).

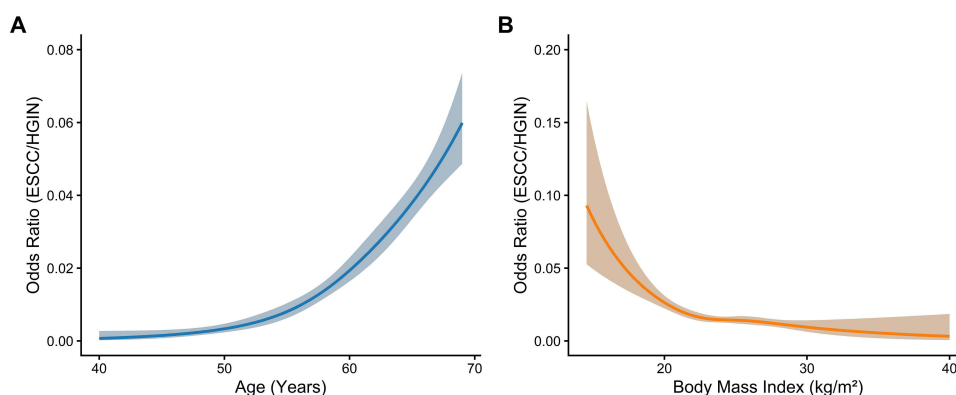


Figure 1 Restricted cubic spline (RCS) analysis of continuous variables. **(A)** RCS curve smoothing the non-linear relationship between age and the risk of ESCC/HGIN. **(B)** RCS curve smoothing the non-linear relationship between body mass index (BMI) and the risk of ESCC/HGIN. The solid lines represent the estimated odds ratios, and the shaded areas represent the 95% confidence intervals.

The 3,938 participants were subsequently divided into a training set and a validation set at an 8:2 ratio using stratified sampling. The training comprised 3,150 individuals, including 284 positive cases, while the validation set included 788 individuals, of whom 74 were positive cases. Baseline characteristics were compared between the two groups, and no statistically significant differences were observed across any variables (all $P > 0.05$), indicating good comparability and supporting their use for model development and internal split-sample validation, respectively (Table 2).

Table 1 Baseline Characteristics of the 3,938 Participants After Random Sampling

Variable	Missing Rate (%)	Lesions Below HGIN (n = 3, 580)	HGIN or Above (n = 358)	P-value
Age (years)	0	55.752±7.848	62.684±5.457	<0.001
BMI (kg/m ²)	6.2	24.302±3.022	23.37±3.188	<0.001
Sex (n[%])	0			<0.001
Male		1794 (50.11)	301 (84.08)	
Female		1786 (49.89)	57 (15.92)	
Educational level (n[%])	0			0.001
Below senior high school		2475 (69.13)	279 (77.93)	
Senior high school or above		1105 (30.87)	79 (22.07)	
Smoking history (n[%])	0			<0.001
No		3138 (87.65)	247 (68.99)	
Yes		442 (12.35)	111 (31.01)	
Alcohol consumption (n[%])	0			0.003
No		757 (21.15)	101 (28.21)	
Yes		2823 (78.85)	257 (71.79)	
Tea drinking habits	0			
No		1263 (35.28)	113 (31.56)	0.178
Yes		2317 (64.72)	245 (68.44)	

(Continued)

Table 1 (Continued).

Variable	Missing Rate (%)	Lesions Below HGIN (n = 3, 580)	HGIN or Above (n = 358)	P-value
Fresh vegetable intake	0			0.027
Never		23 (0.64)	2 (0.56)	
Occasionally (< twice per week)		3407 (95.17)	330 (92.18)	
Frequently ($\geq$ twice per week)		150 (4.19)	26 (7.26)	
Fresh fruit intake	0			0.002
Never		48 (1.34)	7 (1.96)	
Occasionally (< twice per week)		2475 (69.13)	215 (60.06)	
Frequently ($\geq$ twice per week)		1057 (29.53)	136 (37.99)	
Meat, egg, and dairy product intake	0			0.112
Never		14 (0.39)	2 (0.56)	
Occasionally (< twice per week)		3138 (87.65)	300 (83.80)	
Frequently ($\geq$ twice per week)		428 (11.96)	56 (15.64)	
Legume intake	0			0.439
Never		42 (1.17)	7 (1.96)	
Occasionally (< twice per week)		2102 (58.72)	207 (57.82)	
Frequently ($\geq$ twice per week)		1436 (40.11)	144 (40.22)	
Pickled food intake	0			0.009
Never		426 (11.90)	37 (10.34)	
Occasionally (< twice per week)		557 (15.56)	78 (21.79)	
Frequently ($\geq$ twice per week)		2597 (72.54)	243 (67.88)	
Fried food intake	0			<0.001
Never		339 (9.47)	32 (8.94)	
Occasionally (< twice per week)		235 (6.56)	44 (12.29)	
Frequently ($\geq$ twice per week)		3006 (83.97)	282 (78.77)	
Very hot food/beverage consumption	0			<0.001
Never		1112 (31.06)	87 (24.30)	
Occasionally (< twice per week)		430 (12.01)	66 (18.44)	
Frequently ($\geq$ twice per week)		2038 (56.93)	205 (57.26)	
Mold-contaminated food intake	0			0.002
Never		3156 (88.16)	301 (84.08)	
Occasionally (< twice per week)		23 (0.64)	8 (2.23)	
Frequently ($\geq$ twice per week)		401 (11.20)	49 (13.69)	

(Continued)

Table 1 (Continued).

Variable	Missing Rate (%)	Lesions Below HGIN (n = 3, 580)	HGIN or Above (n = 358)	P-value
History of gastrointestinal disease	0			0.325
No		2250 (62.85)	215 (60.06)	
Yes		1330 (37.15)	143 (39.94)	
Family history of esophageal cancer				0.051
No		3197 (89.30)	307 (85.75)	
Yes		383 (10.70)	51 (14.25)	

Table 2 Comparison of Baseline Characteristics Between the Training and Validation Sets

Variable	Training Set (n = 3,150)	Validation Set (n = 788)	P-value
Endoscopic pathology			0.796
Below HGIN	2866 (90.98)	714 (90.61)	
HGIN or above	284 (9.02)	74 (9.39)	
Age (years)	56.315±7.938	56.652±7.826	0.707
BMI (kg/m ²)	23.921±3.219	23.75±3.266	0.076
Sex (n[%])			0.089
Male	1654 (52.51)	441 (55.96)	
Female	1496 (47.49)	347 (44.04)	
Educational level (n[%])			0.577
Below senior high school	2196 (69.71)	558 (70.81)	
Senior high school or above	954 (30.29)	230 (29.19)	
Smoking history (n[%])			0.069
No	2724 (86.48)	661 (83.88)	
Yes	426 (13.52)	127 (16.12)	
Alcohol consumption (n[%])			0.937
No	685 (21.75)	173 (21.95)	
Yes	2465 (78.25)	615 (78.05)	
Tea drinking habits			0.513
No	1109 (35.21)	267 (33.88)	
Yes	2041 (64.79)	521 (66.12)	
Fresh vegetable intake			0.654
Never	20 (0.63)	5 (0.63)	
Occasionally (< twice per week)	2994 (95.05)	743 (94.29)	
Frequently (≥ twice per week)	136 (4.32)	40 (5.08)	
Fresh fruit intake			0.798
Never	44 (1.40)	11 (1.40)	
Occasionally (< twice per week)	2144 (68.06)	546 (69.29)	
Frequently (≥ twice per week)	962 (30.54)	231 (29.31)	
Meat, egg, and dairy product intake			0.986
Never	13 (0.41)	3 (0.38)	
Occasionally (< twice per week)	2749 (87.27)	689 (87.44)	
Frequently (≥ twice per week)	388 (12.32)	96 (12.18)	

(Continued)

Table 2 (Continued).

Variable	Training Set (n = 3,150)	Validation Set (n = 788)	P-value
Legume intake			0.248
Never	36 (1.14)	13 (1.65)	
Occasionally (< twice per week)	1834 (58.22)	475 (60.28)	
Frequently ($\geq$ twice per week)	1280 (40.63)	300 (38.07)	
Pickled food intake			0.859
Never	370 (11.75)	93 (11.80)	
Occasionally (< twice per week)	503 (15.97)	132 (16.75)	
Frequently ($\geq$ twice per week)	2277 (72.29)	563 (71.45)	
Fried food intake			0.415
Never	305 (9.68)	66 (8.38)	
Occasionally (< twice per week)	218 (6.92)	61 (7.74)	
Frequently ($\geq$ twice per week)	2627 (83.40)	661 (83.88)	
Very hot food/beverage consumption			0.913
Never	964 (30.60)	235 (29.82)	
Occasionally (< twice per week)	396 (12.57)	100 (12.69)	
Frequently ($\geq$ twice per week)	1790 (56.83)	453 (57.49)	
Mold-contaminated food intake			0.112
Never	2759 (87.59)	698 (88.58)	
Occasionally (< twice per week)	21 (0.67)	10 (1.27)	
Frequently ($\geq$ twice per week)	370 (11.75)	80 (10.15)	
History of gastrointestinal disease			0.422
No	1982 (62.92)	483 (61.29)	
Yes	1168 (37.08)	305 (38.71)	
Family history of esophageal cancer			0.554
No	2808 (89.14)	696 (88.32)	
Yes	342 (10.86)	92 (11.68)	

Factors Associated with HGIN or HGIN+

The occurrence of HGIN or HGIN+ was defined as the dependent variable. Variable preselection was performed in the training set (n = 3,150) using LASSO regression. The optimal penalty parameter (λ) was determined via 10-fold cross-validation. According to the minimum criterion ($\lambda_{\min}$), 14 candidate predictors were identified, including sex, age, educational level, BMI, smoking history, alcohol consumption, fresh fruit intake, intake of meat/eggs/dairy products, mold-contaminated food consumption, fried food intake, consumption of very hot food or beverages, legume intake, history of gastrointestinal disease, and family history of esophageal cancer (Figure 2).

Before final model construction, collinearity diagnostics were performed for the seven retained predictors (*ie.*, age, sex, educational level, BMI, smoking history, consumption of very hot food or beverages, and family history of esophageal cancer).

These 14 variables were subsequently entered into a multivariate logistic regression model. A backward stepwise procedure was applied, with a removal criterion of $P > 0.1$. Given the established epidemiological relevance of family history of esophageal cancer, the retention threshold was set at $P < 0.1$ to avoid excluding clinically meaningful predictors. Ultimately, seven variables were retained in the final risk model. The results indicated that female sex, educational attainment of senior high school or above, and higher BMI were independently associated with a lower risk of HGIN+. In contrast, older age, smoking, occasional consumption of very hot food or beverages (< twice per week), and a family history of esophageal cancer were identified as significant risk factors. See Table 3.

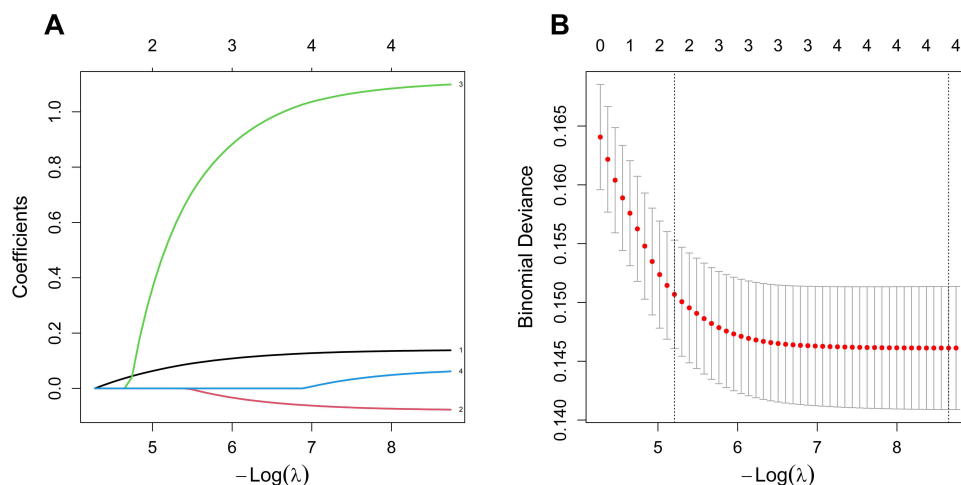


Figure 2 Feature selection using the LASSO regression model. **(A)** LASSO coefficient profiles of the potential predictors against the $\log(\lambda)$ sequence. **(B)** Tuning parameter (λ) selection via 10-fold cross-validation based on the minimum criterion ($\lambda_{\min}$). The vertical dashed lines indicate the optimal λ value chosen.

Clinical Utility and Cost-Effectiveness Evaluation

Before final model construction, collinearity diagnostics were performed for the final set of seven independent predictors (ie, age, sex, educational level, BMI, smoking history, consumption of very hot food or beverages, and family history of esophageal cancer). The results confirmed that all VIFs ranged from 1.01 to 1.16, well below the conventional threshold of 5, indicating no meaningful multicollinearity among the included variables in the final multivariable model. These findings support the stability and interpretability of the regression coefficients. See [Table 4](#).

Table 3 Multivariate Logistic Regression Analysis for HGIN or HGIN+

Variable	Regression Coefficient (β)	Standard Error	OR	95% CI	P-value
Sex					
Male	-	-	1.000	(Reference)	
Female	-1.547	0.183	0.213	0.147 ~ 0.303	<0.0001
Age (years)	0.150	0.012	1.162	1.135~1.190	<0.0001
Educational level					
Below senior high school	-	-	1.000	(Reference)	
Senior high school or above	-0.375	0.168	0.687	0.494~0.956	0.026
BMI	-0.087	0.023	0.917	0.877~0.959	0.0001
Smoking history					
No	-	-	1.000	(Reference)	
Yes	0.525	0.167	1.690	1.218~2.345	0.002
Alcohol consumption					
No	-	-	1.000	(Reference)	
Yes	-0.147	0.161	0.864	0.630~1.184	0.362
Fresh fruit intake					
Never	-	-	1.000	(Reference)	
Occasionally (< twice per week)	-0.181	0.496	0.834	0.316~2.199	0.715
Frequently ($\geq$ twice per week)	-0.019	0.494	0.981	0.373~2.584	0.969
Meat, egg, and dairy product intake					
Never	-	-	1.000	(Reference)	
Occasionally (< twice per week)	1.057	1.191	2.877	0.279~29.620	0.375
Frequently ($\geq$ twice per week)	1.283	1.195	3.608	0.347~35.525	0.283

(Continued)

Table 3 (Continued).

Variable	Regression Coefficient (β)	Standard Error	OR	95% CI	P-value
Mold-contaminated food intake					
Never	-	-	1.000	(Reference)	
Occasionally (< twice per week)	0.074	0.308	1.076	0.588~1.968	0.810
Frequently ($\geq$ twice per week)	0.214	0.269	1.239	0.731~2.100	0.425
Fried food intake					
Never	-	-	1.000	(Reference)	
Occasionally (< twice per week)	0.482	0.3614	1.619	0.797~3.287	0.182
Frequently ($\geq$ twice per week)	-0.155	0.276	0.857	0.500~1.471	0.575
Very hot food/beverage consumption	0.501				
Never	-	-	1.000	(Reference)	
Occasionally (< twice per week)	0.501	0.235	1.652	1.041~2.621	0.032
Frequently ($\geq$ twice per week)	0.277	0.173	1.319	0.940~1.851	0.111
Legume intake					
Never	-	-	1.000	(Reference)	
Occasionally (< twice per week)	-0.592	0.621	0.553	0.164~1.868	0.340
Frequently ($\geq$ twice per week)	-0.627	0.618	0.534	0.159~1.793	0.310
History of gastrointestinal disease					
No	-	-	1.000	(Reference)	
Yes	-0.073	0.144	0.930	0.701~1.233	0.612
Family history of esophageal cancer					
No	-	-	1.000	(Reference)	
Yes	0.361	0.205	1.435	0.960~2.145	0.077

Table 4 Collinearity Diagnostics for Variables Included in the Multivariate Logistic Regression Model

Variable	VIF	Tolerance (= 1/VIF)
Sex	1.12	0.89
Age (years)	1.16	0.86
Educational level	1.08	0.93
BMI	1.05	0.95
Smoking history	1.14	0.88
Very hot food/beverage consumption	1.01	0.99
Family history of esophageal cancer	1.03	0.97

Development and Validation of the Esophageal Cancer Risk Prediction Model

Based on the seven independent predictors for esophageal cancer, a multivariate logistic regression model was established to identify high-risk individuals. The model equation was defined as:

$$\text{Logit}(P) = -8.731 - 1.547 \times X_1 + 0.150 \times X_2 - 0.375 \times X_3 - 0.087 \times X_4 + 0.525 \times X_5 + 0.501 \times X_6 + 0.361 \times X_7$$

(Note: X_1 = sex [female = 1, male = 0]; X_2 = age [years]; X_3 = educational level [senior high school or above = 1, below senior high school = 0]; X_4 = BMI [kg/m^2]; X_5 = smoking history [yes = 1, no = 0]; X_6 = occasional consumption of very hot food or beverages [< twice per week = 1, never = 0]; X_7 = family history of esophageal cancer [yes = 1, no = 0]).

A nomogram was constructed to facilitate individualized risk estimation (Figure 3). Each predictor was assigned a weighted score proportional to its regression coefficient. For instance, a 65-year-old man (approximately 83 points for age plus 34 points for sex), with a history of smoking (approximately 12 points) and a BMI of 20 kg/m^2 (approximately

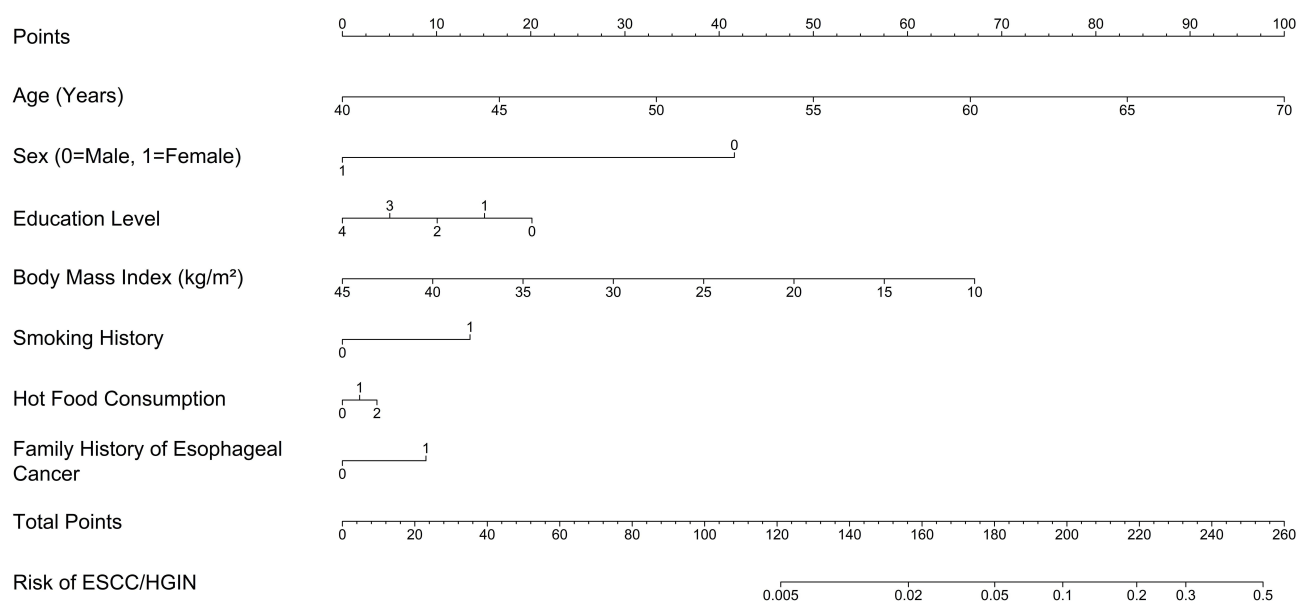


Figure 3 A simplified nomogram for primary population-based screening of esophageal cancer. Each predictor (age, sex, education level, BMI, smoking history, hot-food consumption, and family history of esophageal cancer) is assigned points based on its clinical stratification. The descriptive axis labels and risk scales have been optimized to ensure intuitive, reader-friendly visualization for clinicians.

29 points), would accumulate a total score of roughly 158 points. This corresponds to an estimated probability of 45.3% (0.453) for HGIN or HGIN+. The nomogram thus enables rapid and intuitive visualization of individualized risk.

The AUC was 0.823 in the training set and 0.835 in the validation set, indicating good discriminative capacity for distinguishing high-risk individuals from those at lower risk (Figure 4).

Calibration plots demonstrated close agreement between predicted and observed risks in both the training and validation cohorts, suggesting satisfactory model calibration (Figure 5).

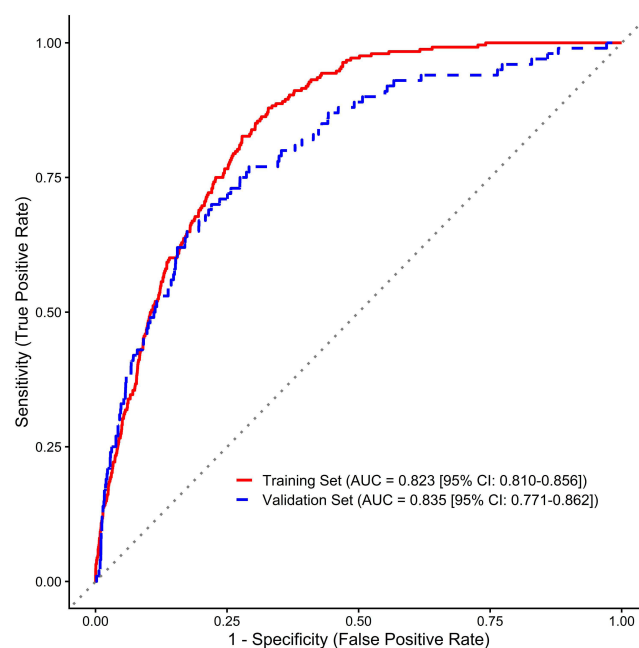


Figure 4 Receiver operating characteristic (ROC) curves of the prediction model. The ROC curves illustrate the discriminative performance of the risk prediction model. The red solid line represents the performance in the training set (AUC = 0.823, 95% CI: 0.810–0.856), and the blue dashed line represents the performance in the validation set (AUC = 0.835, 95% CI: 0.771–0.862). The exact AUC and confidence interval values are prominently integrated into the figure panel.

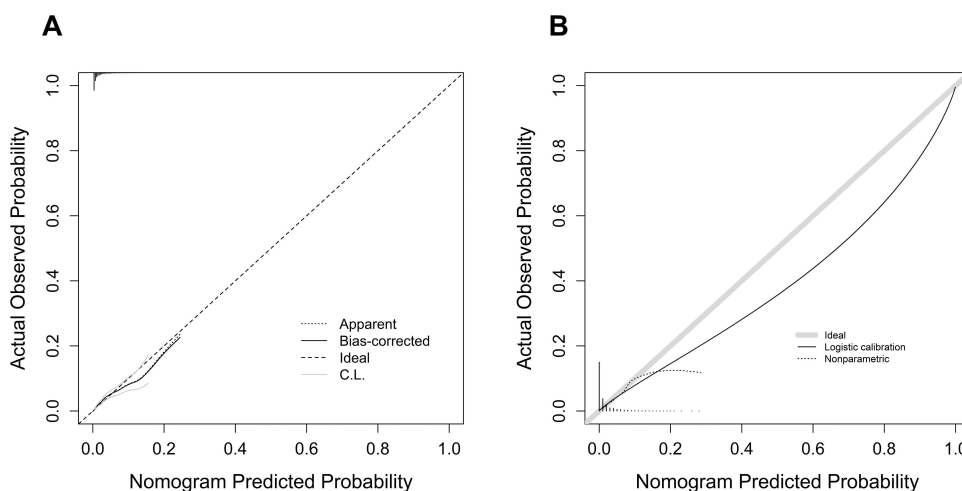


Figure 5 Calibration curves for the risk prediction model. Comparison of predicted versus observed risk in the **(A)** Training set and **(B)** Internal validation set, based on 1,000 bootstrap resamples. Statistical text overlays have been removed from the panels to reduce visual density and emphasize the alignment with the ideal diagonal reference line.

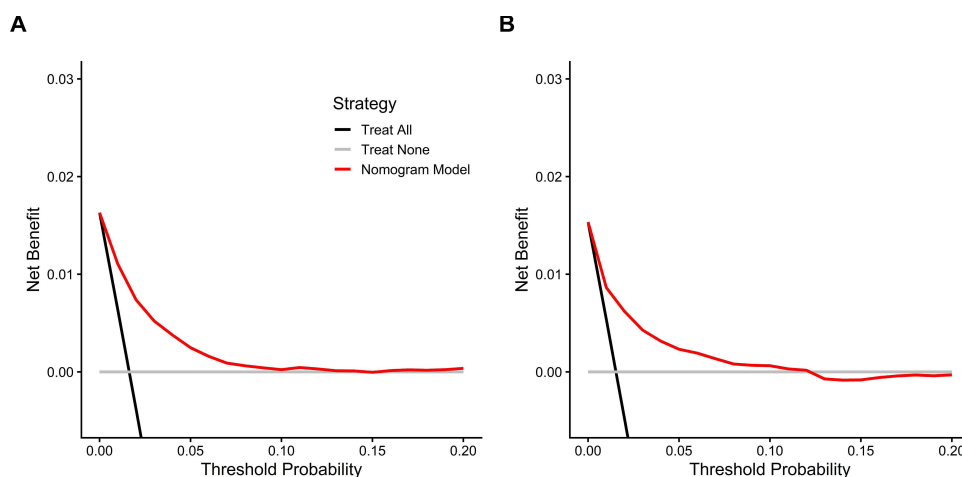


Figure 6 Decision curve analysis (DCA) to assess clinical utility. Net benefit curves for the **(A)** Training set and **(B)** Internal validation set across a range of threshold probabilities. The decision curves have been simplified by removing background grids to reduce visual density, and strategy labels have been updated to clearly distinguish the “Nomogram Model” from alternative default strategies.

DCA was conducted to quantify the clinical utility of the model across a range of threshold probabilities. In the training set, the model consistently yielded a higher net benefit than either the “treat-all” or “treat-none” strategies when the threshold probability ranged from 0 to 0.6. In the validation set, a stable and clinically meaningful net benefit was observed within the threshold interval of 0–0.3, with the most pronounced advantage in the lower end of this range. These findings indicate that the model demonstrates favorable clinical utility under low-threshold conditions, which are commonly adopted in primary screening settings (Figure 6 and Table 5).

According to the discrimination, calibration, and DCA results, the proposed nomogram exhibits robust overall performance and holds promise for practical application.

Using Youden’s index, the optimal cutoff probability was determined to be 0.0678. Participants were accordingly classified into a high-risk group (predicted probability ≥ 0.0678) and a low-risk group (predicted probability < 0.0678). To evaluate practical impact, this cutoff was applied retrospectively to the entire eligible population ($n = 21,717$) for cost-benefit analysis. The results suggest substantial potential for resource optimization (Figure 7). First, the low-risk

Table 5 Net Benefit of the Prediction Model in the Validation Set

Threshold Probability	Model Net Benefit	Default Strategy Net Benefit (Treat All/None)	Net Benefit Difference
0.05	0.0627	0.0427 (All)	0.0201
0.1	0.0427	0.0000 (None)	0.0427
0.15	0.0342	0.0000 (None)	0.0342
0.2	0.0267	0.0000 (None)	0.0267
0.25	0.0123	0.0000 (None)	0.0123
0.3	0.0014	0.0000 (None)	0.0014

group accounted for 63.34% of the total population. In the real-world scenario, restricting endoscopic screening to the high-risk group would therefore reduce endoscopic examinations, as well as the associated personnel and equipment demands, by approximately 63.34%, substantially alleviating the burden on primary healthcare facilities. Second, the high-risk group comprised only 36.66% of the screened population yet captured 87.31% of all detected cases of esophageal cancer and its precursor lesions. This indicates that the model can identify most clinically significant lesions while substantially reducing unnecessary procedures. Collectively, these findings are hypothesis-supporting and operationally promising for enhancing screening efficiency and promoting a more targeted, resource-conscious approach to the early detection of esophageal cancer, although they do not yet establish implementation readiness in broader screening practice.

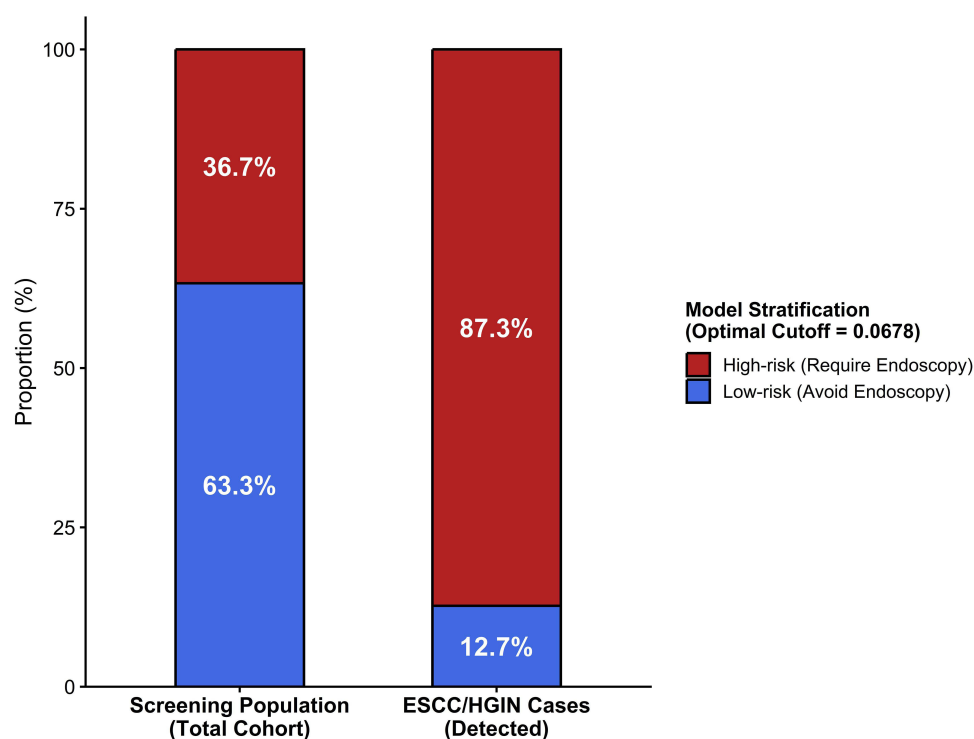


Figure 7 Clinical utility and potential reduction in screening volume. Using the optimal cutoff probability of 0.0678 identified by Youden's index, the population was stratified into high-risk and low-risk groups. The bar chart features explicit risk group segmentations ("Avoid Endoscopy" vs "Require Endoscopy") to clearly communicate that 63.3% of the participants were categorized as low-risk, suggesting that over 60% of unnecessary endoscopic examinations could potentially be avoided. The remaining 36.7% categorized as high-risk captured 87.31% of all ESCC/HGIN cases.

Discussion

Esophageal carcinogenesis typically follows a stepwise progression from low-grade intraepithelial neoplasia to HGIN, and ultimately to invasive carcinoma. Given that HGIN is a critical and potentially reversible stage, timely identification and intervention at this juncture can substantially reduce the likelihood of malignant transformation.⁴ In contrast to most previous prediction models that have used ESCC as the sole endpoint,^{10,11} we defined a composite outcome including HGIN and HGIN+. This design was intended to better align the model with the objectives of early detection. The resulting model integrates readily obtainable clinical variables, including demographic characteristics, lifestyle factors, and family history. It demonstrated robust discriminatory capacity in both the training and validation cohorts, with AUC values of 0.823 (95% confidence interval [95% CI]: 0.798–0.848) and 0.835 (95% CI: 0.805–0.865), respectively, along with satisfactory calibration and favorable clinical utility. This framework is consistent with the “early screening and early treatment” principle advocated in the China Guideline for the Screening of *Esophageal Cancer*⁴ and offers a pragmatic tool for risk stratification in population-based screening, adding practical value for resource-limited implementation.

When developing a model intended for implementation at the primary care level, the selection of objective and easily ascertainable predictors is essential.¹² Some previously reported models, although exhibiting strong statistical performance, are constrained by their reliance on high-cost variables. For example, the model proposed by Yang et al¹³ achieved an AUC of 0.88, indicating excellent discrimination; however, several of its predictors pose practical challenges in large-scale community screening. Accurate quantification of tooth loss typically requires examination by trained dental professionals, increasing logistical complexity and personnel costs in field surveys. Similarly, assessment of the ALDH2 gene expression entails molecular testing, which is associated with substantial expense and demands specialized laboratory infrastructure, technical expertise, and controlled sample transport conditions. In the context of uneven healthcare resource distribution and limited public health funding in many regions of China, models incorporating such high-cost and technically demanding variables are difficult to implement widely. By comparison, the present model is based exclusively on age, basic anthropometric measures, and lifestyle-related variables that can be collected efficiently through structured questionnaires, without the need for laboratory assays or endoscopic evaluation. While maintaining satisfactory predictive performance, this approach markedly lowers operational barriers and enhances feasibility in community settings. Although the model developed by Han et al⁸ demonstrated outstanding discrimination (C-statistic = 0.91), it relied on endoscopic lesion characteristics and therefore cannot be applied before endoscopic screening. In contrast, the seven variables included in our model are accessible through the questionnaire survey, making it particularly suitable for large-scale primary screening where cost containment and operational simplicity are paramount.

Our findings indicate that female sex, higher educational levels, and elevated BMI functioned as protective factors, whereas advanced age, smoking, consumption of very hot food or beverages, and a family history of esophageal cancer were associated with increased risk. These results are broadly consistent with previous reports. Age, as a central predictor, has an effect size comparable to that described by Dag Holmberg et al,¹⁴ who observed an approximate 2% increase in esophageal cancer risk with each additional year of age. In this study, female sex was inversely associated with the risk of ESCC, in agreement with Wang et al,¹⁵ who documented a markedly higher incidence among men. Supporting this sex disparity, Yang et al¹⁶ reported that 59.2% of esophageal cancer deaths among Chinese men were attributable to smoking, highlighting tobacco exposure as a major contributor to the elevated male burden. Both smoking and habitual intake of very hot food or beverages were identified as significant risk factors in our model. The biological plausibility of these associations is consistent with the “thermal injury-smoking synergy” hypothesis proposed by Yu et al,¹⁷ whereby repeated thermal irritation combined with tobacco-related carcinogens leads to chronic cycles of mucosal inflammation and repair, thereby accelerating dysplastic progression. An unexpected pattern was observed with respect to thermal exposure: occasional consumption of very hot food was significantly associated with risk (OR = 1.65), whereas frequent consumption did not reach statistical significance ($P = 0.11$). This discrepancy may reflect competing mechanisms. Chronic thermal exposure might induce mucosal hyperkeratosis, conferring partial tolerance to repeated injury. Alternatively, recall bias could have led to an underestimation of exposure. Because these findings are not entirely intuitive, they should be interpreted carefully. Future investigations using intraoral thermal sensors may help quantify exposure more precisely and clarify dose-response relationships. The protective effects of higher BMI and educational

levels are in line with the findings of Vingeliene et al¹⁸ and are further supported by Shi et al¹². These associations may operate through at least two pathways. First, higher educational levels are often accompanied by improved health literacy and dietary patterns; our data corroborate this inference (Table 1), as the proportion of frequent fresh vegetable consumption was 37.2% in the higher-education group, compared with 28.1% among those with lower educational attainment. Second, education was positively associated with smoking cessation (OR = 1.69), potentially interrupting the synergistic pathway linking hot beverage consumption and tobacco exposure described by Yu et al¹⁷. In addition, Chen et al¹⁹ demonstrated a strong association between family history of esophageal cancer and ESCC risk, a finding consistent with our results. Although family history showed borderline statistical significance in our multivariate model ($P = 0.077$), we retained it due to its well-established epidemiological and clinical relevance, though we acknowledge this warrants careful interpretation and further validation. Collectively, these predictors include both non-modifiable characteristics (eg, age, sex, family history) and modifiable behavioral factors (eg, smoking and dietary habits). Their integration into a single model not only ensures robust predictive power and effective guidance but also provides actionable targets for individualized risk stratification and precision prevention strategies.

At present, most prediction models for esophageal cancer risk are developed using logistic or Cox proportional hazards regression. However, considerable heterogeneity exists in variable selection strategies, data preprocessing, and validation procedures. With respect to continuous predictors such as age and BMI, many studies have transformed them into categorical variables, as seen in the study of Liu et al²⁰ and Chen Yu et al²¹. Although convenient, such categorization may entail loss of information, potentially compromising model performance. Shi et al¹² retained age and BMI as continuous variables yet did not formally assess the assumption of linearity between these predictors and the outcome. In this study, we evaluated the functional form of age and BMI using RCS, with four knots placed at the 5th, 35th, 65th, and 95th percentiles. After confirming approximate linear relationships with the outcome, both variables were incorporated into the model in continuous form. This approach adheres more closely to statistical best practices, preserves the granularity of the original data, and enables more accurate estimation of effect sizes, thereby enhancing predictive precision and practical reliability. Regarding variable selection, Chen et al²¹ applied a conventional sequential strategy, namely, univariate screening followed by multivariate regression, without fully addressing potential multicollinearity among predictors. In contrast, this study employed LASSO regression in combination with multivariate logistic regression, thereby mitigating collinearity and minimizing model bias. In terms of data imbalance, similar to other studies based on gastrointestinal cancer screening datasets,^{10,11,22} the initial prevalence of positive outcomes in our cohort was low (approximately 1.54%). The serious class imbalance can bias model training toward the majority class. Drawing on the strategy proposed by Fu et al¹⁰ for rare complications, a 1:10 case-control sampling scheme was implemented to recalibrate the dataset before modeling. This study did not adopt the commonly used 1:4 ratio because, under an event rate as low as 1.54%, that configuration would have yielded a relatively small sample, potentially widening confidence intervals for a model incorporating seven predictors and the assessment of associations between continuous variables. By selecting a 1:10 ratio, we maximized the informational contribution of the non-event population while preserving risk ranking capacity. The resulting model achieved stable and improved discrimination in the validation cohort (AUC = 0.835), with satisfactory calibration and clinical utility. Although case-control sampling may inflate absolute risk estimates—a factor that must be carefully considered when interpreting its clinical applicability for absolute risk communication—it does not affect risk ordering, which plays a central role in stratified screening. DCA further demonstrated significant net benefit across a threshold probability range of 0–0.6. Using the optimal cutoff determined by the Youden index (0.0678), risk stratification revealed that the high-risk group comprised 36.66% of the overall population yet accounted for 87.31% of detected cases. In practical terms, based on model-based estimates from our dataset, application of this model could potentially avert approximately 63.34% of unnecessary endoscopic examinations, thereby theoretically improving allocation efficiency of limited endoscopic resources in primary care settings. However, given the retrospective design, the resampled 1:10 case-control nature of the data, and current lack of external validation, these findings represent potential utility for risk stratification rather than established practice-level outcomes. Therefore, these estimates of clinical impact and the nomogram-derived absolute risk probabilities should be interpreted cautiously, and evidence sufficient for real-world implementation is not yet established.

Several limitations should be acknowledged. First, this was a single-center study, and all participants were recruited from Tai'an, Shandong Province, an area with a high incidence of esophageal cancer. The regional concentration of the sample introduces geographic specificity. Variations in dietary patterns (eg, consumption of pickled foods), environmental exposures, and genetic background across different regions may influence disease risk profiles. Accordingly, the extrapolability and temporal stability of this model to low-incidence settings or populations with distinct demographic and environmental characteristics requires confirmation through multicenter external validation. Second, the retrospective design relies primarily on baseline questionnaire data to identify exposure variables. Third, while the 1:10 case-control resampling strategy effectively handled class imbalance for risk ranking, it inevitably alters the baseline prevalence. This artificial inflation of the event rate may affect absolute risk estimation and calibration when applied to the unselected target population, thereby limiting the transportability of the model beyond the present cohort until properly recalibrated. Self-reported information on long-term habits (eg, frequency of consuming very hot food, cumulative smoking exposure) may be subject to imperfect recall, thereby introducing recall bias. In addition, given that the dataset was derived from an opportunistic screening program, a degree of selection bias cannot be excluded. Individuals who voluntarily participate in screening may differ systematically from the general population, potentially exhibiting greater health awareness or a stronger family history of disease. As a result, risk estimates generated by the model could be inflated if applied indiscriminately to the broader population. Future research may advance along several directions. First, prospective, multicenter studies are needed to further validate and optimize the model. Second, incorporating machine learning approaches may help capture complex nonlinear interactions among predictors and enhance predictive power. Third, translating the model into user-friendly digital tools (eg, mobile applications, web-based calculators) would promote its integration into primary care workflows and support clinical decision-making at the community level.

To facilitate real-time risk assessment by clinicians and public health practitioners, we developed an online risk calculator using the R Shiny framework. By entering a subject's baseline characteristics, users can obtain an individualized predicted probability and corresponding risk stratification based on the optimal cutoff identified in this study (0.0678).

Conclusion

In summary, while similar prediction models already exist, this study adds practical value by explicitly integrating class-imbalance correction techniques into a simplified primary screening model for esophageal cancer. The methodological framework provides a replicable approach for handling rare-event imbalance in similar cancer screening cohorts. The resulting model, comprising only seven readily obtainable variables, demonstrates satisfactory discrimination and calibration while remaining operationally simple. It is applicable to individuals aged 40–69 years without a prior history of malignancy in Tai'an, Shandong Province. By enabling efficient identification of high-risk individuals, the model provides a scientific basis for optimizing endoscopic resource allocation and holds potential value for population-level prevention strategies as a risk-stratification tool. Nevertheless, given the retrospective, single-center design and reliance on internal validation, these conclusions must be framed cautiously. Specifically, absolute risk estimates and clinical cutoff-based recommendations should be interpreted with care due to the resampled nature of the data, and the model requires further confirmation through prospective external validation before broader implementation.

Data Sharing Statement

All data generated or analyzed during this study are included in the article.

Ethics Approval and Consent to Participate

This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Tai'an Cancer Hospital (Tai'an Cancer Center, Tai'an Third People's Hospital) (Approval No: 232016).

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically

reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

2024 Tai'an City Science and Technology Innovation and Development Projects in the Fields of Agriculture and Social Development (Policy-Guided Category); Development and Visualization of an Esophageal Cancer Incidence Prediction Model Based on Logistic Regression and Machine Learning Algorithms (No. 2024NS158).

Disclosure

None of the authors have any personal, financial, commercial, or academic conflicts of interest in this work.

References

1. Han B, Zheng R, Zeng H, et al. Cancer incidence and mortality in China, 2022. *J Natl Cancer Cent.* 2024;4(1):47–53. doi:10.1016/j.jncc.2024.01.006
2. 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. *CA Cancer J Clin.* 2024;74(3):229–263. doi:10.3322/caac.21834
3. Zeng H, Chen W, Zheng R, et al. Changing cancer survival in China during 2003–15: a pooled analysis of 17 population-based cancer registries. *Lancet Glob Health.* 2018;6(5):e555–e567. doi:10.1016/S2214-109X(18)30127-X
4. He J, Chen WQ, Li ZS, et al. China guideline for the screening, early detection and early treatment of esophageal cancer (2022, Beijing). *Zhonghua Zhong Liu Za Zhi.* 2022;44(6):491–522. doi:10.3760/cma.j.cn112152-20220517-00348
5. Wei WQ, Chen ZF, He YT, et al. Long-Term follow-up of a community assignment, one-time endoscopic screening study of esophageal cancer in China. *J Clin Oncol.* 2015;33(17):1951–1957. doi:10.1200/JCO.2014.58.0423
6. Chen Q, Yu L, Hao C, et al. Effectiveness evaluation of organized screening for esophageal cancer: a case-control study in Linzhou city, China. *Sci Rep.* 2016;6:35707. doi:10.1038/srep35707
7. Hippisley-Cox J, Mei W, Fitzgerald R, et al. Development and validation of a novel risk prediction algorithm to estimate 10-year risk of oesophageal cancer in primary care: prospective cohort study and evaluation of performance against two other risk prediction models. *Lancet Reg Health Eur.* 2023;32:100700. doi:10.1016/j.lanepe.2023.100700
8. Han J, Guo X, Zhao L, et al. Development and validation of esophageal squamous cell carcinoma risk prediction models based on an endoscopic screening program. *JAMA Network Open.* 2023;6(1):e2253148. doi:10.1001/jamanetworkopen
9. Han Y, Zhu X, Hu Y, et al. Electronic health record-based absolute risk prediction model for esophageal cancer in the chinese population: model development and external validation. *JMIR Public Health Surveill.* 2023;9:e43725. doi:10.2196/43725
10. Fu Y, Li KB, Yang XC. A risk score model for predicting cardiac rupture after acute myocardial infarction. *Chin Med J.* 2019;132(9):1037–1044. doi:10.1097/CM9.000000000000175
11. Chen W, Li H, Ren J, et al. Selection of high-risk individuals for esophageal cancer screening: a prediction model of esophageal squamous cell carcinoma based on a multicenter screening cohort in rural China. *Int J Cancer.* 2021;148(2):329–339. doi:10.1002/ijc.33208
12. Shi DQ, Li YY, Hao CQ, et al. Optimization of initial screening strategies for esophageal cancer in high-incidence areas based on predictive models. *Chin J Cancer Prevent Treat.* 2025;32(9):524–531. doi:10.16073/j.cnki.cjcp.2025.09.02
13. Yang X, Suo C, Zhang T, et al. A nomogram for screening esophageal squamous cell carcinoma based on environmental risk factors in a high-incidence area of China: a population-based case-control study. *BMC Cancer.* 2021;21(1):343. doi:10.1186/s12885-021-08053-7
14. Holmberg D, Ness-Jensen E, Mattsson F, et al. Clinical prediction model for tumor progression in Barrett's esophagus. *Surg Endosc.* 2019;33(9):2901–2908. doi:10.1007/s00464-018-6590-5
15. Wang QL, Ness-Jensen E, Santoni G, et al. Development and validation of a risk prediction model for esophageal squamous cell carcinoma using cohort studies. *Am J Gastroenterol.* 2021;116(4):683–691. doi:10.14309/ajg.0000000000001094
16. Yang H, Fan JH, Qiao YL. Analysis of disease burden and attributable risk factors of esophageal cancer in countries with high incidence of esophageal cancer from 1990 to 2019. *Cancer Research and Clinic.* 2023;35(10):721–727. doi:10.3760/cma.j.cn115355-20221215-00791
17. Yu C, Tang H, Guo Y, et al. Hot Tea consumption and its interactions with alcohol and tobacco use on the risk for esophageal cancer: a Population-Based Cohort Study. *Ann Intern Med.* 2018;168(7):489–497. doi:10.7326/M17-2000
18. Vingeliene S, Chan DSM, Vieira AR, et al. An update of the WCRF/AICR systematic literature review and meta-analysis on dietary and anthropometric factors and esophageal cancer risk. *Ann Oncol.* 2017;28(10):2409–2419. doi:10.1093/annonc/mdx338
19. Chen T, Cheng H, Chen X, et al. Family history of esophageal cancer increases the risk of esophageal squamous cell carcinoma. *Sci Rep.* 2015;5:16038. doi:10.1038/srep16038
20. Liu M, Zhou R, Liu Z, et al. Update and validation of a diagnostic model to identify prevalent malignant lesions in esophagus in general population. *EClinicalMedicine.* 2022;47:101394. doi:10.1016/j.eclim.2022.101394
21. Chen Y, Zhang H, Xu K, et al. A predictive model for high-risk population of esophageal squamous cell carcinoma and precancerous lesions based on upper gastrointestinal cancer screening programme in rural areas of Shandong. *China Cancer.* 2021;30(11):820–826.
22. Liu M, Liu Z, Cai H, et al. A model to identify individuals at high risk for esophageal squamous cell carcinoma and precancerous lesions in regions of high prevalence in China. *Clin Gastroenterol Hepatol.* 2017;15(10):1538–1546.e7. doi:10.1016/j.cgh.2017.03.019

Risk Management and Healthcare Policy**Dovepress**
Taylor & Francis Group**Publish your work in this journal**

Risk Management and Healthcare Policy is an international, peer-reviewed, open access journal focusing on all aspects of public health, policy, and preventative measures to promote good health and improve morbidity and mortality in the population. The journal welcomes submitted papers covering original research, basic science, clinical & epidemiological studies, reviews and evaluations, guidelines, expert opinion and commentary, case reports and extended reports. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/risk-management-and-healthcare-policy-journal>