

Prevalence of *Helicobacter pylori* infection among young adults aged 18–25 years and development of machine learning-based screening models

Yu Zhang¹

Xinjiang 474 Hospital

Xin Peng

First Affiliated Hospital of Shihezi University Medical College

Miaomiao Luo

First Affiliated Hospital of Shihezi University Medical College

Haowei Li

First Affiliated Hospital of Shihezi University Medical College

Guoming Luo

First Affiliated Hospital of Shihezi University Medical College

Minghan Zhang

First Affiliated Hospital of Shihezi University Medical College

Hao Liu

First Affiliated Hospital of Shihezi University Medical College

Weigang Chen

13579456959@126.com

First Affiliated Hospital of Shihezi University Medical College

Research Article

Keywords: Young adults aged 18–25 years, *Helicobacter pylori*, Screening model, Model performance, Machine learnin

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10937543/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

1 **Prevalence of *Helicobacter pylori* infection among young adults aged**
2 **18–25 years and development of machine learning-based screening**
3 **models**

4 Yu Zhang¹, Xin Peng², Miaomiao Luo², Haowei Li², Guoming Luo²,
5 Minghan Zhang², Hao Liu², Weigang Chen^{2*}

6 Youth Digestive Cooperative Group, Gastroenterology Branch of Chinese
7 Medical Association

8 ¹Department of Gastroenterology, Xinjiang 474 Hospital, Xinjiang, China

9 ²The First Affiliated Hospital, Shihezi University School, Xinjiang, China

10 ORCID

11 Yu Zhang :<https://orcid.org/0000-0002-5430-6772>

12 **Correspondence:** Weigang Chen, Department of Gastroenterology, The
13 First Affiliated Hospital, Shihezi University School of Medicine, Xinjiang
14 832002, China. Email: 13579456959@126.com

15 **Running head:** *Helicobacter pylori* infection Screening among Young
16 Adults

Abstract

Background: *Helicobacter pylori* (*H. pylori*) is a major modifiable risk factor for benign and malignant gastroduodenal diseases. Young adults aged 18–25 years represent an optimal window for *H. pylori* eradication with high cost-effectiveness. However, epidemiological evidence and targeted screening tools for *H. pylori* infection in this specific age group remain limited. This study aimed to investigate the prevalence of *H. pylori* infection among people aged 18–25 years, and to construct and validate machine learning-based screening models for population-wide *H. pylori* risk assessment.

Materials and Methods: A random cluster sampling method was adopted to recruit participants aged 18–25 years from Shihezi University. All eligible participants underwent serological testing, *H. pylori* stool antigen test and ¹³C-urea breath test, and completed a structured electronic questionnaire. The overall dataset was randomly divided into a training set and a validation set at a ratio of 7:3. Three analytical approaches, including logistic regression, extreme gradient boosting (XGB) and support vector machine (SVM), were applied to establish screening models. The receiver operating characteristic (ROC) curve and corresponding area under the curve (AUC), sensitivity, specificity, accuracy, precision and F1-score were used to comprehensively evaluate the performance of each model. Variable importance was further analysed

based on the optimal model.

Results: A total of 832 valid participants were enrolled in the final analysis, including 583 in the training set and 249 in the validation set. The SVM model exhibited the best predictive performance, with a sensitivity of 70.3%, a specificity of 77.4%, an accuracy of 0.753, a precision of 0.565 and an AUC of 0.835. According to the variable importance ranking of the SVM model, the top five influencing factors were family size, prior H. pylori testing history, annual household income, gastroscopy history and household pet ownership, respectively. The overall H. pylori infection rate among the 18–25-year-old population was 29.1%.

Conclusions: H. pylori infection is prevalent among young adults aged 18–25 years in this study setting. The SVM-based screening model achieves satisfactory predictive efficacy and can serve as a convenient and efficient tool for large-scale H. pylori infection screening in this young population.

Keywords: Young adults aged 18–25 years; Helicobacter pylori; Screening model; Model performance; Machine learnin

1. Introduction

Helicobacter pylori (*H. pylori*) infection is recognised as the predominant modifiable risk factor for a spectrum of benign and malignant gastroduodenal disorders worldwide. The overall *H. pylori* infection rate in the general Chinese population ranges from 40% to 60%. Accumulated evidence has demonstrated that population-based *H. pylori* screening and eradication can reduce gastric cancer incidence by approximately 40%.

Screening and eradication of *H. pylori* in young adults yield substantial public health benefits, including the relief of *H. pylori*-associated gastritis, reduction of gastric cancer risk, and interruption of intra-familial transmission of *H. pylori*. There is ongoing debate regarding the appropriate age for initiating routine *H. pylori* screening and intervention.

The Kyoto Global Consensus Report recommended *H. pylori* screening and eradication for adolescents aged 12–18 years to reduce gastric cancer burden and improve economic returns. Notably, previous research indicated that *H. pylori* eradication among individuals aged 18–25 years presents the highest cost-effectiveness, as gastric mucosal atrophy and intestinal metaplasia rarely occur in this age group. Nevertheless, large-scale epidemiological investigations and dedicated screening tools for *H. pylori* infection among 18–25-year-olds are still scarce.

As a branch of artificial intelligence, machine learning techniques are capable of learning patterns from complex datasets and supporting

79 automated decision-making, which have been widely applied in medical
80 data management, analysis and risk prediction. Several studies have
81 suggested that predictive models constructed using individual and
82 household characteristics can effectively identify populations at high risk
83 of *H. pylori* infection, showing promising application prospects in public
84 health practice.

85 Given the above research gaps, this cross-sectional study was conducted
86 to clarify the epidemiological characteristics of *H. pylori* infection among
87 young adults aged 18–25 years. We further developed and validated
88 multiple machine learning models to identify key risk factors and
89 establish low-cost, practical screening tools for early *H. pylori* detection
90 in this population.

91 2. Subjects and Methods

92 2.1 Study participants

93 This cross-sectional study was carried out among individuals aged 18–25
94 years from Shihezi University. Based on sample size calculation, an
95 estimated 843 participants were required. A total of 870 individuals
96 initially participated in the survey and laboratory tests, and 832 valid
97 datasets were finally included for analysis, corresponding to a
98 questionnaire response rate of 95.6%.

99 This study was approved by the Ethics Committee of The First Affiliated
100 Hospital of Shihezi University (Approval No.: KJ2024-471-01). All

participants provided electronic informed consent prior to enrolment. All procedures performed in this study were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments.

Inclusion criteria: (1) Aged between 18 and 25 years; (2) Voluntary participation and signed informed consent.

Exclusion criteria: (1) Individuals with pre-existing cardiovascular, hepatic, renal or other chronic systemic diseases; (2) Participants with mental disorders; (3) Those who failed to meet the requirements for serological testing, H. pylori stool antigen test or ¹³C-urea breath test; (4) Participants who withdrew from the study during the investigation.

According to the results of three laboratory tests, participants with at least two positive test results were classified into the H. pylori-infected group (n=242), and the remaining participants were assigned to the non-infected group (n=590).

2.2 Data collection and outcome definition

The electronic questionnaire used in this study was adapted from a previously validated questionnaire applied in a national multicentre epidemiological survey of H. pylori infection led by Professor Zhaoshen Li. Additional items concerning willingness to receive H. pylori screening and influencing factors of screening adherence were supplemented after expert discussion. The questionnaire maintained good reliability and

validity for data collection.

H. pylori infection status was determined by three non-invasive tests, namely serological assay, *H. pylori* stool antigen test (*H. pylori*SAT) and ¹³C-urea breath test (¹³C-UBT). A definitive positive diagnosis of *H. pylori* infection was defined as two or more positive results across the three tests.

2.3 Statistical analysis

SPSS 27.0 and R 4.3.2 software were used for statistical analysis. Continuous variables with normal distribution were presented as mean $\pm$ standard deviation and compared using the independent samples t-test. Non-normally distributed continuous variables were described as median (interquartile range, IQR) and analysed via the Mann-Whitney U test. Categorical variables were expressed as counts and percentages, and compared using the Chi-square test or Fisher's exact test where appropriate. A two-tailed P value < 0.05 was considered statistically significant.

For variable selection in multivariable analysis, univariate analysis was first performed, and variables with $P < 0.150$ were retained to avoid omitting potential influential factors. Selected variables were then incorporated into binary logistic regression analysis.

For machine learning model construction: a 10-fold cross-validation strategy was used for XGB model training, and variable importance was

evaluated based on gain values. A 5-fold cross-validation was adopted for SVM model training, and variable importance was determined by the absolute value of weight coefficients. The ROC curve and AUC were used to evaluate the discriminative ability of each model. Sensitivity, specificity, accuracy, precision and F1-score were calculated to assess the overall predictive performance. Accuracy was defined as the proportion of correctly classified samples among all participants, and precision referred to the proportion of true positives in all predicted positives. The optimal model was selected for subsequent visualisation and practical application analysis.

3. Results

3.1 Prevalence of *H. pylori* infection

Among the 832 enrolled participants, 242 were *H. pylori*-positive, including 112 males and 130 females, with a median age of 21 (19, 24) years. A total of 590 participants were *H. pylori*-negative, consisting of 266 males and 324 females, with a median age of 21 (20, 23) years. No significant differences in age, gender or ethnicity were observed between the two groups (all $P > 0.05$) (Table 1).

3.2 Binary logistic regression model

After univariate analysis, six variables including family size, annual household income, gastrointestinal symptoms, gastroscopy history, prior *H. pylori* testing history and gastroduodenal surgery history ($P < 0.150$,

Table 2) were included in the binary logistic regression model. The results indicated that family size (OR = 1.82, P = 0.005) and prior H. pylori testing history (OR = 0.45, P = 0.008) were independent influencing factors for H. pylori infection (Table 3).

For the logistic regression model, the training set yielded an AUC of 0.690 (95% CI: 0.64–0.74), a sensitivity of 17.4%, a specificity of 95.3%, an accuracy of 0.724, a precision of 0.612 and an F1-score of 0.271. The validation set presented an AUC of 0.693 (95% CI: 0.63–0.76), a sensitivity of 15.7%, a specificity of 94.4%, an accuracy of 0.722, a precision of 0.524 and an F1-score of 0.242. Figure 1. The logistic regression model showed moderate and stable discriminative capacity, but its low sensitivity limited the comprehensive performance for balancing positive detection and predictive reliability.

3.3 XGB model

Variable importance analysis of the XGB model showed that the top five influential factors were family size, prior H. pylori testing history, age, residential area size and maternal occupation. Figure 2a. In the XGB model, the training set had an AUC of 0.757 (95% CI: 0.71–0.80), a sensitivity of 73.3%, a specificity of 61.8%, an accuracy of 0.652, a precision of 0.445 and an F1-score of 0.554. The validation set obtained an AUC of 0.722 (95% CI: 0.65–0.79), a sensitivity of 77.1%, a specificity of 62.6%, an accuracy of 0.667, a precision of 0.446 and an

F1-score of 0.565. Figure2b. The XGB model had moderate discriminative performance and could be applied for preliminary screening of high-risk populations, while clinical re-evaluation was recommended due to its relatively low precision.

3.4 SVM model

According to the variable importance ranking of the SVM model, the top five predictors were family size, prior H. pylori testing history, annual household income, gastroscopy history and household pet ownership. Figure3a. The SVM model achieved an AUC of 0.835 (95% CI: 0.80–0.87), a sensitivity of 70.3%, a specificity of 77.4%, an accuracy of 0.753 and a precision of 0.565 in the training set. The validation set showed an AUC of 0.708 (95% CI: 0.63–0.78), a sensitivity of 55.7%, a specificity of 73.2%, an accuracy of 0.683, a precision of 0.448 and an F1-score of 0.497. Figure3b. The SVM model presented strong discriminative ability and was identified as the optimal screening model. Further optimisation with larger sample sizes is required to improve its generalisation performance.

3.5 Performance comparison of the three models

Based on the indicators from the training set (Table 4), the AUC values ranked as SVM > logistic regression > XGB; sensitivity ranked as XGB > SVM > logistic regression; specificity ranked as logistic regression > SVM > XGB; accuracy ranked as SVM > logistic regression > XGB; and

F1-score ranked as SVM > XGB > logistic regression. The logistic regression model had extremely low sensitivity and was only suitable for scenarios requiring high specificity. The XGB model had high sensitivity but relatively low precision, which may lead to a high false-positive rate. Overall, the SVM model had the best discriminative ability and balanced comprehensive performance, making it the most suitable tool for H. pylori infection preliminary screening.

3.6 Visualisation of the SVM decision model

The visualised decision graph of the SVM model was established. Figure 4. Samples with high classification confidence were marked in red, and those with low confidence were marked in blue. Three groups of populations with high H. pylori infection risk were prioritised for screening: participants from families with more than three members and without prior H. pylori testing; participants from households with annual income below 100,000 yuan and a family size larger than three; participants from low-income households (annual income < 100,000 yuan) without prior H. pylori testing.

4. Discussion

H. pylori infection is a leading aetiological factor for chronic active gastritis, peptic ulcer disease and gastric cancer, with a global overall infection rate of approximately 50%. As the most important modifiable risk factor for gastric cancer, early identification and intervention of H.

pylori infection are critical for public health and disease prevention. Previous studies confirmed that *H. pylori* eradication among people aged 18–25 years yields the highest cost-effectiveness, since gastric mucosal atrophy and intestinal metaplasia rarely occur in this age group, forming a key time window for gastric cancer prevention. Nevertheless, systematic epidemiological research and dedicated screening models for *H. pylori* infection in this young adult population remain insufficient. This study constructed machine learning-based screening models using easily accessible individual and household indicators, aiming to improve screening efficiency and reduce economic costs for population-based *H. pylori* prevention and control.

In the present study, the overall *H. pylori* infection rate among 18–25-year-old young adults was 29.1%, which was lower than the reported infection rate (40%–60%) in the general Chinese population. This finding may be attributed to higher educational attainment and better public health awareness among university students. However, the infection rate in our cohort was higher than the previously reported rate of 24.1% among Chinese college students and 25.8% among young adults in South Korea. Although young adults have a relatively lower *H. pylori* burden compared with the general population, targeted screening and health education are still necessary.

Multivariable analysis revealed that a family size of more than three

members was an independent risk factor for *H. pylori* infection (OR = 1.82), which strongly indicated obvious intra-familial clustering of *H. pylori* infection, consistent with the known oral-oral and faecal-oral transmission routes of *H. pylori*. Participants with prior *H. pylori* testing history had a lower infection risk (OR = 0.45), which may be related to better health awareness or previous successful *H. pylori* eradication treatment. Annual household income was also confirmed as an important influencing factor by the SVM model, in line with previous findings that low household income is associated with elevated *H. pylori* infection risk. Age was not identified as an independent risk factor in this study, and it may affect *H. pylori* infection through interactions with household environment and socioeconomic factors.

Machine learning algorithms have been increasingly applied to develop predictive models for infectious diseases in recent years. The SVM model established in this study had better predictive performance than previously reported *H. pylori* prediction models. The visualised decision graph further simplified model interpretation and facilitated practical application in public health settings. Family size, prior *H. pylori* testing history and annual household income were verified as the core predictors, which can guide rapid risk stratification and targeted screening among young adults.

It should be noted that the predictive performance of the SVM model

decreased in the independent validation set compared with the training set, suggesting limited generalisation ability due to single-centre sampling. Future studies will focus on feature optimisation, hyperparameter adjustment and external validation across multiple centres and different populations to enhance model robustness. In addition, several potential influencing factors, such as dietary habits, oral health status and *H. pylori* virulent strains, were not included in this questionnaire-based survey, which can be supplemented in further research to improve model efficacy. In conclusion, this study clarified the epidemiological characteristics and key risk factors of *H. pylori* infection among young adults aged 18–25 years. The established SVM-based screening model has good predictive performance and practical value for low-cost, large-scale *H. pylori* screening. This work provides a useful tool for population-based *H. pylori* prevention and contributes to primary prevention of gastric cancer. Further external validation and model optimisation in broader populations are required in follow-up research.

Authors' contributions

YZ designed the study, performed statistical analysis, drafted and revised the manuscript. MML, XP and HL participated in health education and clinical consultation. HWL, GML and MHZ were responsible for data collection and sorting. WGC conceived the overall research scheme, managed research funds, supervised the whole study and revised the final

manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by the 2024 Talent Development Fund (Grant No. CZ001228), the Key Laboratory for Prevention and Treatment of High Prevalent Diseases in Central Asia (Co-construction) of National Health Commission (Grant No. 2020-PT330-003), and the Bingtuan Guiding Science and Technology Program (Grant No. 2023ZD009).

References

1. Jiyan L, Yingying H, Jialun G, Biyu H, Caihong W, Fangyi S, Peiyuan L: **Survey of willingness to detect *Helicobacter pylori* infection among physical examination population.** *Chinese Journal of Gastroenterology and Hepatology* 2024, **33**(05):502-508. (in Chinese)
2. Zhou X-Z, Lyu N-H, Zhu H-Y, Cai Q-C, Kong X-Y, Xie P, Zhou L-Y, Ding S-Z, Li Z-S, Du Y-Q: **Large-scale, national, family-based epidemiological study on *Helicobacter pylori* infection in China: the time to change practice for related disease prevention.** *Gut*, 2023, **72**(5): 855-869.
3. Liou J-M, Malfertheiner P, Lee Y-C, Sheu B-S, Sugano K, Cheng

- 321 H-C, Yeoh K-G, Hsu P-I, Goh K-L, Mahachai V et al: **Screening and**
322 **eradication of Helicobacter pylori for gastric cancer prevention:**
323 **the Taipei global consensus.** *Gut* 2020, **69**(12):2093-2112.
- 324 4. Sugano K, Tack J, Kuipers EJ, Graham DY, El-Omar EM, Miura S,
325 Haruma K, Asaka M, Uemura N, Malfertheiner P: **Kyoto global**
326 **consensus report on Helicobacter pylori gastritis.** *Gut* 2015,
327 **64**(9):1353-1367.
- 328 5. Li S, Wang Y, Huang R, Yang L, Lyu X, Huang X, Peng X, Song D,
329 Ma N, Xiao Y, Zhou Q, Guo Y, Liang N, Liu S, Gao K. **Development**
330 **of a machine learning-based diagnostic model for T-shaped**
331 **uterus using transvaginal 3D ultrasound quantitative parameters.**
332 *Chinese Medical Journal*, 2025, **105**(30): 2551-2557. (in Chinese)
- 333 6. Haoyue Z, Ling X, Bo X, Shuo H, Yongxiang T, Feng L, Pan T,
334 **Machine learning-based classification of temporal lobe epilepsy**
335 **subtypes and surgical prognosis evaluation using PET metabolic**
336 **networks.** *Chinese Medical Journal*, 2025, **105**(31): 2645-2654.
337 (in Chinese)
- 338 7. Tran V, Saad T, Tesfaye M, Walelign S, Wordofa M, Abera D, Desta
339 K, Tsegaye A, Ay A, Taye B, **Helicobacter pylori (H. pylori) risk**
340 **factor analysis and prevalence prediction: a machine**
341 **learning-based approach.** *BMC Infectious Diseases*, 2022, **22**(1).
- 342 8. Guibin L, Jiaqian G, Lina Z, Jiangtao H, Hongming Z, Yiting L, Yuan

W, Bin C. **Urease Breath Test and Stool Antigen Test Diagnose Helicobacter Pylori Infection in Patients with Bleeding Peptic Ulcer: a Meta-analysis.** *Chinese General Practice*, 2022, **25** (3): 354-362 (in Chinese)

9. Shuai R, Pengpeng C, Yaqian L, Tianpei W, Yan Z, Qian L, Yuanliang G, Liqin W, Caiwang Y, Guangfu J; **Prevalence of Helicobacter pylori infection in China: A systematic review and meta-analysis.** *Journal of gastroenterology and hepatology* 2022 Mar;**37**(3):464-470

10. Hafiz TA, D'Sa JL, Zamzam S, Dionaldo MLV, Aldawood E, Madkhali N, Mubarak MA. **Helicobacter pylori Infection: Comparison of Knowledge between Health Science and Non-Health Science University Students.** *International Journal of Environmental Research and Public Health*, 2021, **18**(15).

11. Zhaoyang W, Ning Z: **A study on the influencing factors of college students'willingness to undergo Helicobacter pylori screening based on the COM-B model.** *Health Development and and Policy Research*. 2024;**27**(2):125-135. (in Chinese)

12. Lee J-H, Choi KD, Jung H-Y, Baik GH, Park JK, Kim SS, Kim B-W, Hong SJ, Lim H, Shin CM, Lee SH, Jeon SW, Kim JH, Choi CW. **Seroprevalence of Helicobacter pylori in Korea: A multicenter, nationwide study conducted in 2015 and 2016.** *Helicobacter*, 2018, **23**(2).

13. Ying W, Yongmei W, Yunran Y, Qingqing Z, Yu L, Da W, Dan Z, Kunpeng N, Lan L, Yule H, Hang C, Minshan H, Le C, Dingyun Y, Lanqing M; **Epidemiological status of family-based Helicobacter pylori infection in Yunnan Province, China.** *BMC Gastroenterology*, 2025, **25**(1). doi: 10.1186/s12876-025-03806-1
14. National National Clinical Research Center for Digestive Diseases (Shanghai), National Gastrointestinal Early Cancer Prevention and Treatment Alliance, Helicobacter pylori and Peptic Ulcer Group, Chinese Society of Gastroenterology, **National Collaborative Group for Helicobacter pylori Research: Chinese consensus on family based Helicobacter pylori infection control and management.** *Chinese Journal of Digestion*. 2021;**41**(4):221-233. (in Chinese)
15. Quancai C, Chunping Z, Yuan Y, Qi F, Yichao F, Yingxia H, Jichang L, Kaiguang Z, Guoliang Y, Liping Y, Nonghua L, Shengsheng Z, Chengxia , Mingquan L, Qi L, Rongzhou L, Jie P, Xiaocui Y, Xuqing Z, Yumei L, Bo L, Ansheng L, Honghui C, Xiuling L, Ping X, Jianfeng Z, Baozhen L, Zhiqiang D, Yiqi D, Zhaoshen L, **Development and validation of a prediction rule for estimating gastric cancer risk in the Chinese high-risk population: a nationwide multicentre study.** *Gut*, 2019, **68**(9): 1576-1587.
16. Jieling D, Yiming Y, Huisi H, Cuihua W, Yuanxiang L.**Data mining to construct the prediction model and prevention strategy**

387 of *Helico-bacter pylori* infection patients. *Chinese Journal of*
388 *Gastroenterology and Hepatology*.2022, 31(9): 992-998.

Table1 Analysis of H. pylori-infection test findings

Items	Negative group (n=590)	Positive group (n=242)	χ^2/Z	P
Age ^b (years, IQR)	21 (20,23)	21 (19,24)	1.10	0.270
Sex ^a				
Male	266 (70.4)	112 (29.6)	0.10	0.075
Female	324 (71.4)	130 (28.6)		
Ethnicity ^a				
Han	535 (71.8)	210 (28.2)	2.79	0.095
Ethnic minorities	55 (63.2)	32 (36.8)		
Education level ^a				
University	485 (71.5)	193 (28.5)	0.68	0.408
Postgraduate	105 (68.2)	49 (31.8)		

Note: ^an(%); ^b M(Q1,Q3);

Table 2 Univariate analysis of characteristics of H. pylori-infection [n(%)]

Item	Negative group (n=411)	Positive group (n=172)	χ^2	P
Sex			0.10	0.075
Male	266 (70.4)	112 (29.6)		
Female	324 (71.4)	130 (28.6)		
Ethnicity			2.79	0.095
Han nationality	535 (71.8)	210 (28.2)		
Ethnic minorities	55 (63.2)	32 (36.8)		
Family size			8.50	0.004
≤3 family members	141 (78.8)	38 (21.2)		
>3 family members	270 (66.8)	134 (33.2)		
Annual household income			2.21	0.027
<100,000 RMB	250 (67.4)	121 (32.6)		
100,000–300,000 RMB	145 (75.5)	47 (24.5)		
>300,000 RMB	16 (80.0)	4 (20.0)		
Presence of gastrointestinal symptoms			2.18	0.140
No	185 (73.7)	66 (26.3)		
Yes	226 (68.1)	106 (31.9)		
History of gastroscopy			2.93	0.087
No	375 (69.6)	164 (30.4)		
Yes	36 (81.8)	8 (18.2)		
Prior Helicobacter pylori testing history			7.39	0.007
No prior test	339 (68.3)	157 (31.7)		
Received H. pylori test before	72 (82.8)	15 (17.2)		
History of gastroduodenal surgery			–	0.079
No	410 (70.8)	169 (29.2)		
Yes	1 (25.0)	3 (75.0)		

Notes: The reference group was the H. pylori-negative group

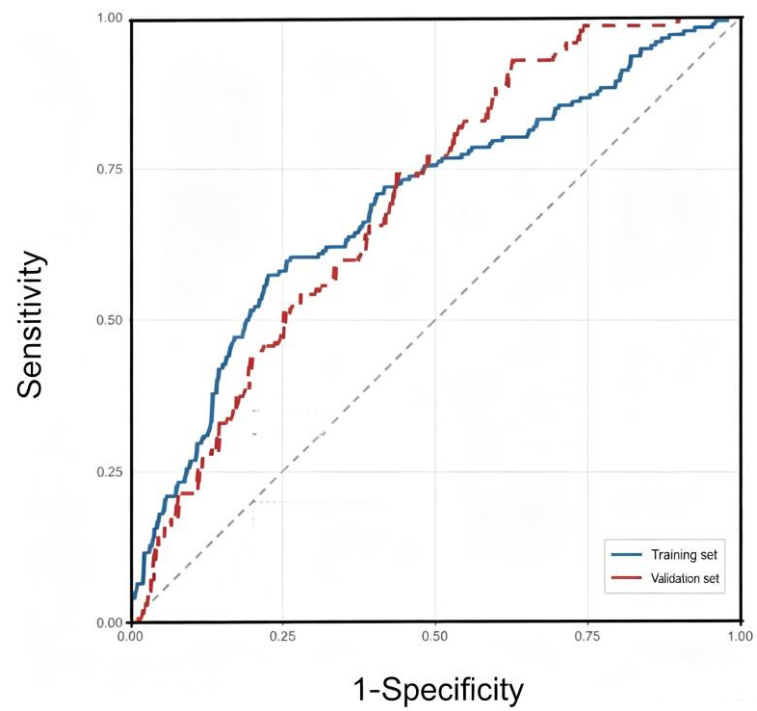
Table 3 Binary Logistic Regression Analysis

Model Step	Included Variables	OR (95%CI)	<i>P</i>
Step 1	Number of family members (1)	1.84 (1.22–2.79)	0.004
	Constant	0.27	<0.001
Step 2	Number of family members (1)	1.82 (1.20–2.76)	0.005
	History of <i>H. pylori</i> testing (1)	0.46 (0.25–0.82)	0.009
	Constant	0.30	<0.001
Step 3	Number of family members (1)	1.82 (1.20–2.77)	0.005
	History of <i>H. pylori</i> testing (1)	0.45 (0.25–0.81)	0.008
	History of gastroduodenal surgery (1)	8.17 (0.82–81.60)	0.074
	Constant	0.30	<0.001

Note: OR= Odds Ratio; *CI*= Confidence Interval

Table 4 Comparison of three models

Method	AUC	Sensitivity	Specificity	Accuracy	Precision	F1-score
Logistic regression	0.690	17.4%	95.3%	0.724	0.612	0.271
XGBoost model	0.757	73.3%	61.8%	0.652	0.445	0.554
SVM model	0.835	70.3%	77.4%	0.753	0.565	0.627



391

Figure1 ROC Curves and AUC of the Logistic Regression Model in the Training and Validation Set

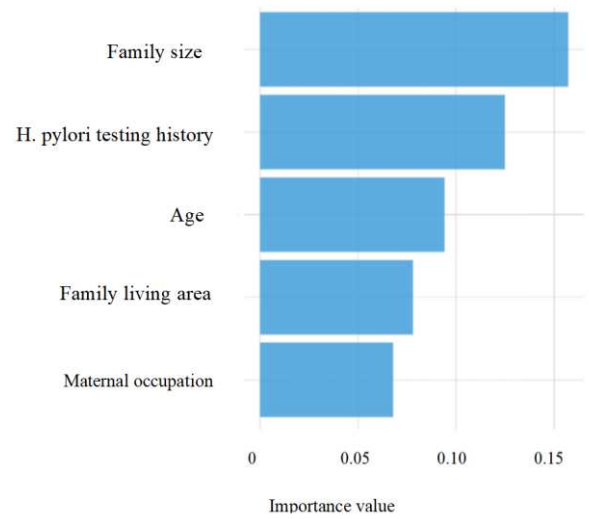


Figure 2a Ranking of Feature Importance for the XGBoost Model

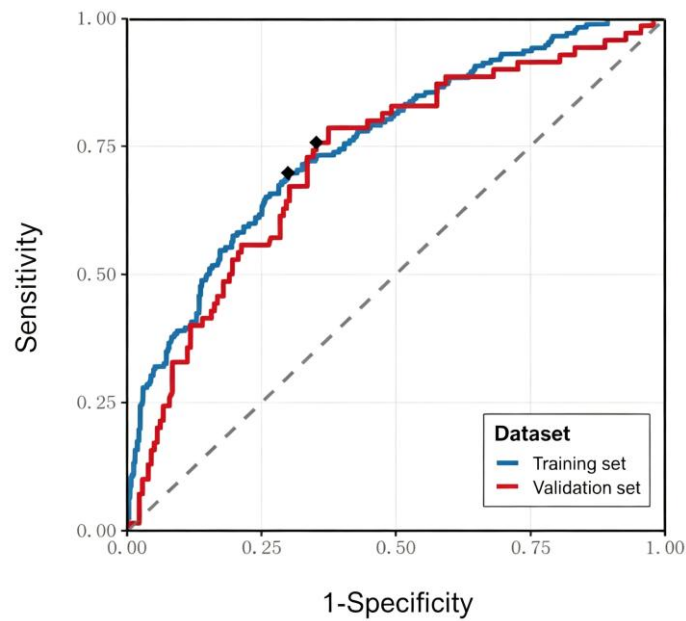


Figure 2b ROC Curves and AUC of the XGBoost Model in the Training and Validation Sets

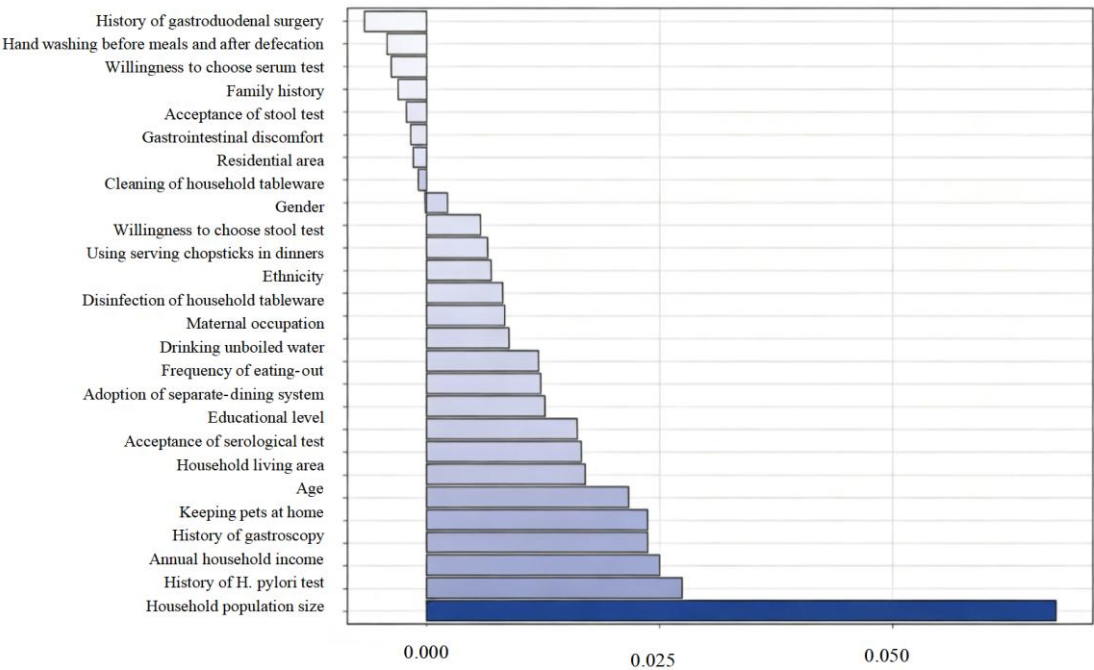


Figure 3a Bar Plot of Feature Importance for the SVM Model

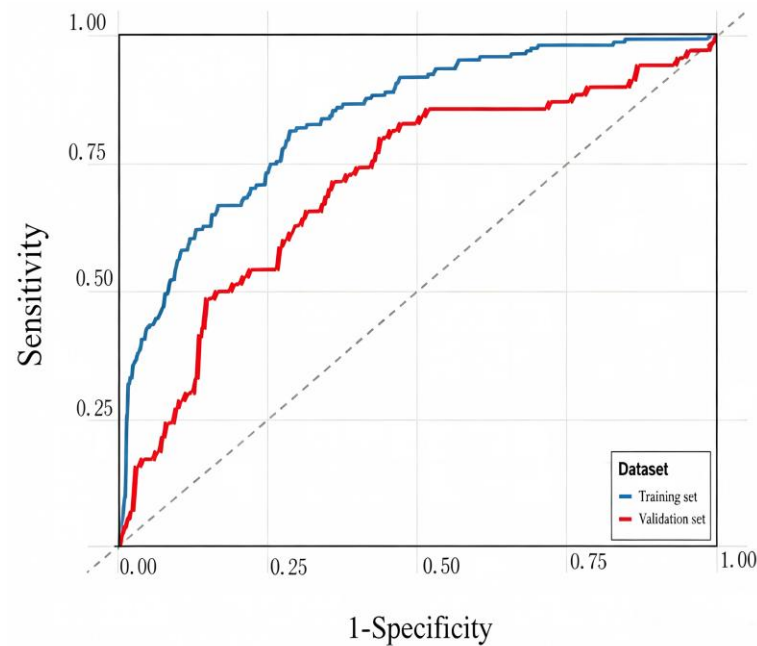


Figure 3b ROC Curves and AUC of the SVM Model in the Training and Validation Sets

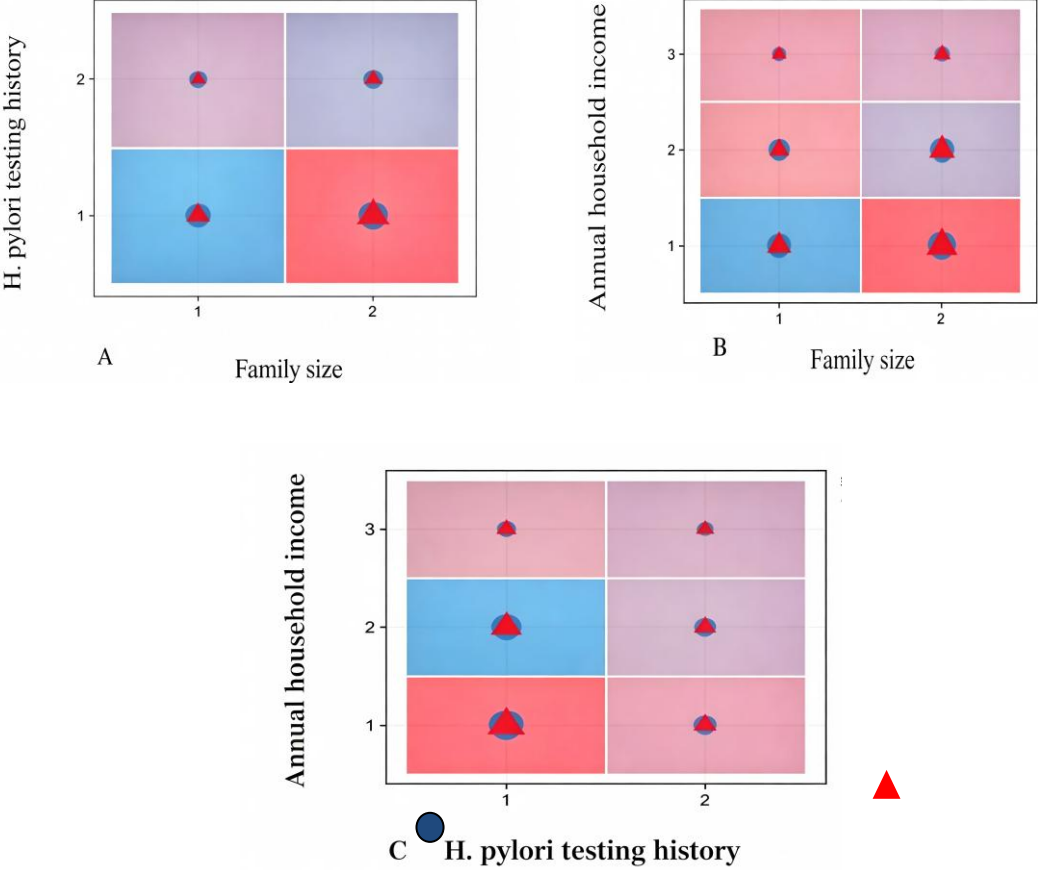


Figure 4 Visualization of the SVM decision boundary
Note:A = Family size vs. Hp test history;B = Family size vs. annual household income;
C = Hp test history vs. annual household income; Outcome measures: Infected,
Non-infected.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Table2.docx](#)
- [Table3.docx](#)
- [Table1.docx](#)
- [Table4.docx](#)