Parasitology Research (2025) 124:26
https://doi.org/10.1007/s00436-025-08470-8

RESEARCH

Epidemiology and risk factors of *Clonorchis sinensis* infection in the mountainous areas of Longsheng County, Guangxi: insights from automated machine learning

Xiaowen Li[1] · Yu Chen[4] · Guoyang Huang[1] · Xuerong Sun[3] · Gang Mo[1] · Xiaohong Peng[1,2]

Received: 30 October 2024 / Accepted: 14 February 2025 / Published online: 5 March 2025

Abstract

Clonorchis sinensis (*C. sinensis*) is mainly prevalent in Northeast and South China, with Guangxi being the most severely affected region. This study aimed to evaluate the prevalence and identify the risk factors of *C. sinensis* infection in Longsheng County, a mountainous area in northern Guangxi. In 2023, a comprehensive study was conducted in Longsheng County, utilizing longstanding inhabitants as study participants. Questionnaires were employed to gather data on fish consumption, awareness of *C. sinensis*, and residential coordinates, while fecal examinations were utilized to identify *C. sinensis* infection. Important risk factors for the *C. sinensis* infection were identified through the development of individual infection risk models using automated machine learning techniques. A total of 740 fecal samples were collected, revealing an overall *C. sinensis* infection rate of 69.59%. The gradient boosting machine (GBM) was the most accurate predictor with an area under the precision-recall Curve (AUPRC) of 0.997. The model identified years of raw fresh fish consumption, frequency of raw fresh fish consumption, elevation, and water distance as the top four predictors of *C. sinensis* infection risk. In conclusion, our study has revealed a high infection rate of *C. sinensis* in the mountainous areas of Longsheng County, with adults, men, and farmers particularly susceptible to both high incidence and infection severity. We developed a high-performance predictive model for individual *C. sinensis* infection within the county, identifying the key risk factors for local infections. These findings offer valuable guidance for the control and prevention of clonorchiasis.

Keywords *Clonorchis sinensis* · Risk factor · Automated machine learning · Guangxi · Prediction

Introduction

Clonorchiasis, also known as infection with the Chinese liver fluke, is a foodborne parasitic infection transmitted through the consumption of raw or inadequately cooked freshwater fish infected with *C. sinensis* metacercaria (Qian et al. 2016; Na et al. 2020). This disease exhibits a high prevalence in various Asian countries and regions, including China, South Korea, Japan, northern Vietnam, and the Russian Far East (Qian et al. 2024). In China, clonorchiasis is a common endemic parasitic disease. Large-scale surveys showed that the average *C. sinensis* infection rate increased from 0.37% between 1988 and 1992 to 0.58% between 2002 and 2004. Although this may seem like a small increase, the number of infected individuals rose from 4.7 million to 12.49 million during this period, indicating a growing trend of the disease nationwide (Lun et al. 2005). According to the 2015 National Survey Report on the Current Status of Major Human Parasitic Diseases, Guangxi emerges as a primary

Section Editor: Pengfei Cai

✉ Gang Mo
958082480@qq.com

✉ Xiaohong Peng
pxh815@163.com

[1] Guangxi University Key Laboratory of Pathogenic Biology, Guilin Medical University, Guilin, Guangxi, China

[2] Guangxi Key Laboratory of Molecular Medicine in Liver Injury and Repair, The Affiliated Hospital of Guilin Medical University, Guilin, Guangxi, China

[3] Academic Affairs Office of Guilin Medical University, Guilin, Guangxi, China

[4] Hengzhou Center for Disease Control and Prevention, Hengzhou, China

endemic region, boasting an urban *C. sinensis* infection rate of 17.48%, ranking it at the forefront nationally (Wan et al. 2019). Moreover, the infection rate of *C. sinensis* in different areas of Guangxi varies greatly, Heng County exhibits the highest prevalence of *C. sinensis* infection at 31.70%, while there are rates approaching zero in non-endemic regions of Guangxi (Lv et al. 2021). This variation may be attributed to disparities in geographical conditions, demographic characteristics, and sociocultural factors across different areas (Kim et al. 2017; Wang et al. 2023). Therefore, conducting a focused analysis of *C. sinensis* risk within a specific, localized area may yield more precise and informative results.

Longsheng County, situated in the northern region of the Guangxi Zhuang Autonomous Region, stands as an autonomous haven for ethnic minorities. This culturally diverse region has a longstanding culinary tradition centered around the consumption of raw fresh fish. The Longsheng area is distinguished by its diverse topography, including numerous mountains, significant elevation variations between population centers and villages, intricate landforms, and intersecting water systems. Many studies have shown that the prevalence of clonorchiasis has obvious regional distribution characteristics, including altitude and distance to water sources (Sripa et al. 2021; Liu et al. 2023). While a local survey of clonorchiasis in Longsheng has been conducted (Rong et al. 2011), it is noteworthy that prior to our study, there was a dearth of systematic research reports examining the risk factors associated with *C. sinensis* infection in this region, particularly social and geographical environmental factors.

Identifying risk factors for *C. sinensis* infection in specific regions is crucial for guiding strategies and intervention programs (Vinh et al. 2017; Lee et al. 2020; Xin et al. 2021). Machine learning (ML) is a scientific discipline that emphasizes efficient computational algorithms and has been widely used in various studies of parasitic diseases in different countries or regions (Roessler et al. 2022; Xu et al. 2023; Li et al. 2024; Zheng et al. 2024). Automated machine learning (AutoML) intelligently evaluates hundreds to thousands of mathematical models, ultimately selecting the optimal one, making it particularly suitable for rapid application by public health professionals with limited computer science expertise (Zhao et al. 2024).

Based on this, our research attempted to predict the risk of *C. sinensis* infection in a small-scale region (Longsheng County) using machine learning, geographic information, and socio-demographic information. Furthermore, this study aims to develop an individual person prediction model for *C. sinensis* infection using automated machine learning techniques. Ultimately, the goal is to predict infection prevalence in targeted areas, thereby facilitating the implementation, evaluation, and sustained effectiveness of intervention strategies.

Materials and methods

Study area

This cross-sectional study was conducted in 2023 in Longsheng County, Guangxi Zhuang Autonomous Region. Nestled in the upper reaches of the Pearl River branch basin, Longsheng County boasts an average altitude hovering between 700 and 800 m, characterized by a rugged and complex natural terrain. The county has a jurisdictional population of approximately 140,000, of which about 64% are farmers. It is a culturally diverse community, comprised of over ten ethnic groups such as the Han, Zhuang, Dong, and Yao.

Research design

Based on the specific geographical features and population distribution of inhabited areas in the study region, it has been divided into three distinct altitude categories: high (700–1200 m), medium (400–699 m), and low (< 400 m). The stratified random sampling technique has been employed to select a total of 10 villages across the three designated strata. According to the formula $n = \frac{z_\alpha \times p(1-p)}{d^2}$ to calculate the sample size (Zhan 2012), the confidence interval is 95%, the estimated infection rate of *C. sinensis* is 60% (Yang et al. 2019), and the tolerance error is $d = 0.1 * P = 0.04$, the calculated sample size was estimated to be 576 people, but considering that 10% of the participants may not be able to provide stool, the sample size increased to 637 people after adjustment. Eligible study participants were required to have resided in the study area for at least 6 months and be at least 3 years old (Xu et al. 2021). GPS technology was utilized to establish the coordinates of survey points and document their respective altitudes. The surveyed area was digitally mapped using ArcGIS10.8 (see Fig. 1). The water distance, expressed in meters (m), represents the Euclidean distance from each grid point to the nearest water source, encompassing lakes, rivers, wetlands, and river beaches.

Questionnaire data

To gather comprehensive demographic data and assess awareness of clonorchiasis among study participants, a standardized questionnaire was employed. This questionnaire encompassed a range of variables, including gender, age, ethnic group, occupation, education level, consumption of raw fresh fish, annual frequency of raw fresh fish consumption, duration of raw fresh fish consumption habits, knowledge of clonorchiasis, and alcohol intake. During the administration of the questionnaire, researchers

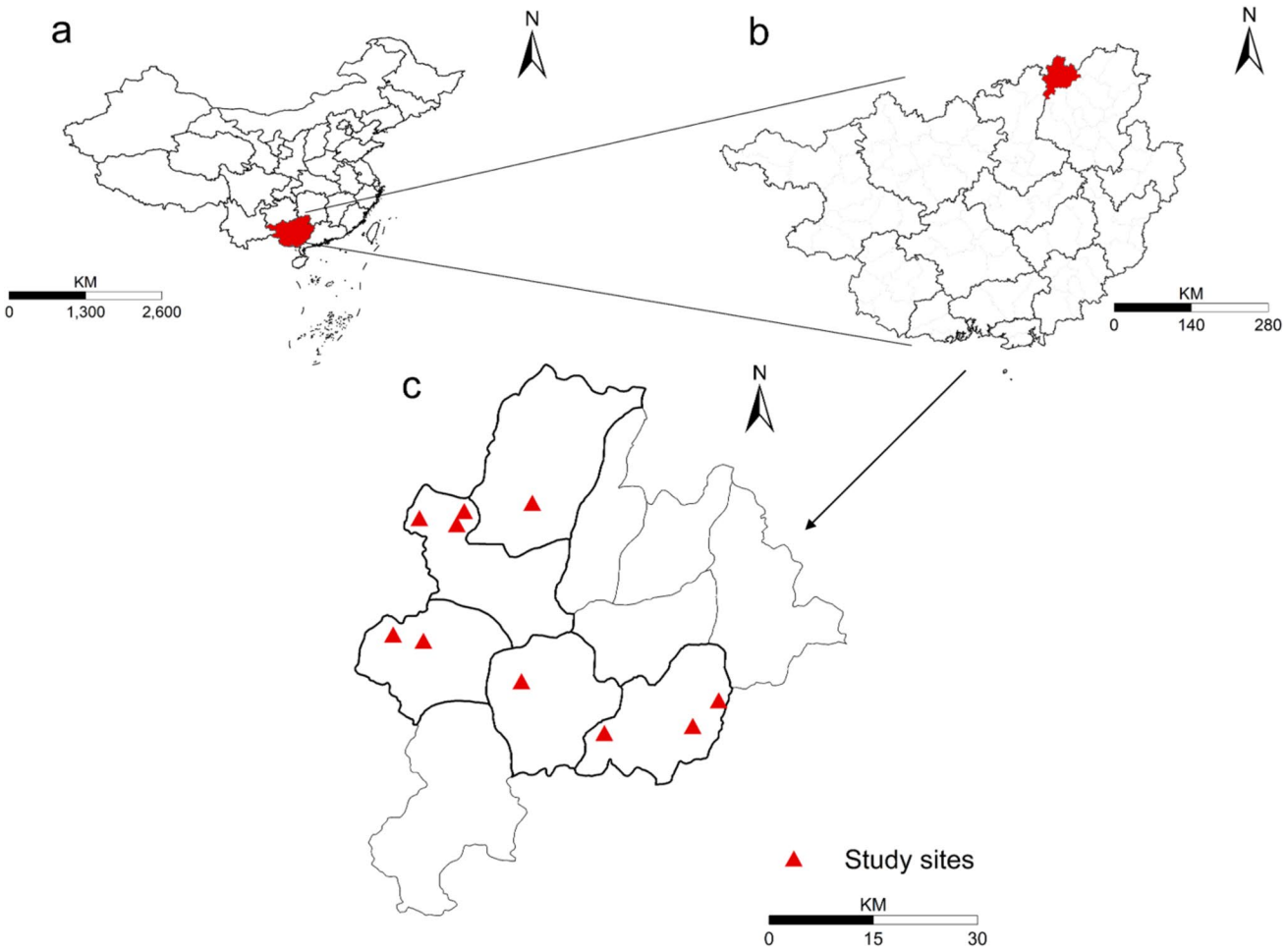

Fig. 1 Geographic location of study sites in Longsheng County, Guangxi Autonomous Region, China

diligently verified and corrected any omitted responses by directly engaging with participants. Consistent quality control measures were implemented throughout the process to ensure the accuracy and reliability of the collected data.

Stool collection and laboratory tests

Each participant was supplied with a labeled plastic sample container for the collection of stool samples, with a minimum weight requirement of 30 g. The modified Kato thick smear method was used to assess both the infection rate and extent of infection with *C. sinensis* (Odongo-Aginya et al. 1995). The prepared Kato-Katz slides were allowed to air-dry at room temperature until transparent. Once dried, the slides were examined under a microscope. The presence of eggs in the smear indicated a positive result, and the number of *C. sinensis* eggs was counted under the microscope, with three smears examined per sample to ensure accuracy.

Definition of infection rate and infection intensity

Utilizing the findings from the investigation into *C. sinensis* infection, a comprehensive analysis was performed using R Studio software (version 4.3.3). The calculation of the proportion of total infections involves dividing the number of infections by the total number of survey participants. The determination of the eggs per gram of feces (EPG) was accomplished by multiplying the number of eggs observed in each smear by 24 and subsequently computing the average of three smears. The infection intensity was categorized into three levels: mild infection (EPG 1–999), moderate infection (EPG 1000–9999), and severe infection (EPG > 10,000) (Fürst et al. 2012).

Data analysis

The data underwent double-entry and cross-validation processes in EpiData 3.1 software (http://www.epidata.dk). The statistical analysis was conducted using the R 4.3.3 software,

and the proportion of mild, moderate, and severe infections was determined by dividing the respective number of infections by the total number of infections. We analyzed and contrasted the rates of infection among various demographic groups, including gender, age, occupation, and education level, as well as the prevalence of three distinct levels of infection severity: mild, moderate, and severe. Age was divided into five categories: under 15 years, 15–29 years, 30–44 years, 45–59 years, and above 59 years (Qian et al. 2011). To comprehend the consumption patterns of raw fresh fish among the study participants, an analysis was conducted to assess the duration of raw fresh fish consumption (0–1 year, 1–5 years, and over 5 years) and the frequency of consumption (0 times per year, once per year, and 2 or more times per year) within the overall population and various subgroups (Sun et al. 2020). The study utilized frequency and percentage to describe the count data and employed either the chi-square test or Fisher's exact probability method to assess variations between groups. A two-tailed test with a significance level of $P < 0.05$ indicated statistical significance in the observed differences.

Model construction based on AutoML algorithms

We selected all 12 variables of this survey to generate a high-performance machine learning model. These variables encompass gender, age, ethnic group, occupation, education level, consumption of raw fresh fish, frequency of raw fresh fish consumption, duration of raw fresh fish consumption, knowledge of *C. sinensis*, alcohol consumption, elevation, and distance to water source. We leveraged the open-source AutoML package for R from the automated machine learning platform, enabling us to download the software locally and analyze survey data without uploading it to external cloud services. H2O.ai AutoML incorporates a range of standard machine learning algorithms for training and cross-validation, including GBM, deep learning (DL), generalized linear model (GLM), distributed randomization forest (DRF), and extremely randomized trees (XRT). Automated machine learning (AutoML) allows public health physicians without coding experience to build their own machine learning models (Milad et al. 2024). In addition, AutoML creates two types of ensemble learning models: one that includes "all" algorithms and another that is a subset of the "best of the family." Detailed information on how each model was built and which hyperparameters were optimized by AutoML can be found in the documentation provided by H2O.ai. Our AutoML models were trained using 80% of the original dataset, which was further divided into five folds for cross-validation. A total of 20 models were developed and ranked based on their AUPRC scores during the cross-validation process. The remaining 20% of the data was reserved as the test set. This test set was used to evaluate the final performance of the top-performing model selected from the cross-validation process. The process cycles through all parts, and the average performance is calculated to evaluate the model's generalization ability. The relevant code for the construction of the model can be found in the public GitHub repository at the following link: (https://github.com/lixiaowen1999/AutoML), which provides detailed instructions for reproducing analysis.

Model interpretability

When the model identifies the key variables within its internal structure, we employ SHAP and PDP plots to further elucidate the functioning of the model (Lundberg et al. 2020). The SHAP chart presents the variables in a hierarchical format, with the most influential ones positioned at the top and the least influential ones resting at the bottom. In this research, the SHAP values, designed as the x-axis, meticulously depict the proportional effect of each variable's value on predicting an individual's likelihood of contracting *C. sinensis* infection. A SHAP value exceeding zero on the x-axis indicates a positive correlation with a higher infection rate, whereas a value below zero suggests a negative correlation with a lower infection rate. Within the SHAP diagram, everyone is represented by a point along the horizontal axis, where each point corresponds to a specific variable. The color of each point signifies the standardized variable values of the patient, adhering to a color-coded distribution, where red denotes high values and blue represents low values.

Additionally, a partial dependence plot (PDP) offers a visual portrayal of the marginal impact of a specific variable on a predicted outcome, such as the infection rate. The mean response value quantifies the variable's influence. In the present investigation, the infection rate exhibited a response value of 1, indicating a 100% probability of infection. PDP plots offer insights into whether the relationship between targets and features is linear, monotonic, or intricate.

Optimizing model with select variables

Afterward, we extracted the four most influential variables based on their significant contributions to the model's performance and retrained the AutoML model using only these variables. This retained model was then evaluated for its performance. We also constructed a model that depended on factors such as altitude, water distance, occupation, ethnicity, alcohol consumption, age, education, gender, and knowledge of *C. sinensis* when the variables related to raw fish consumption were set to zero.

Results

Fundamental attributes of study participants

A total of 740 individuals participated in the study, with 515 confirmed cases of *C. sinensis* infection, translating to an infection rate of 69.59%. Among the infected individuals, 25 (4.85%) were also found to have nematode infection. The infected individuals spanned a broad age range, from 15 to 88 years, with a mean age of 54 years and a median age of 56 years (interquartile range: 34–65 years). Notably, all infected individuals were adults, those aged ≤ 29 years exhibited a significantly lower infection rate compared with older age groups (30.91% vs 70.70%) ($\chi^2 = 40.07$, $P < 0.05$). In terms of occupation and education, 495(96.12%) of the infected individuals were farmers, and 479 (93.01%) had an education level of junior high school or below. Gender-wise, the prevalence of infection was significantly higher among men (75.00%) compared to women (64.10%) ($\chi^2 = 9.40$, $P < 0.05$). Additionally, there was a notable difference between non-farmers (27.40%) and farmers (74.21%) in terms of infection rates ($\chi^2 = 65.96$, $P < 0.05$). There were differences in the infection of *C. sinensis* between individuals living at different altitudes, with the rate of 83.69% at lower altitudes being much higher than 39.81% at middle and high altitudes ($\chi^2 = 226.98$, $P < 0.05$). There were also differences in the infection rate of *C. sinensis* among different ethnic groups ($\chi^2 = 45.72$, $P < 0.05$). There was no significant difference in the infection of *C. sinensis* between individuals who were aware of clonorchiasis and those who were not, with rates of 75.98% and 60.59%, respectively (Table 1).

Infection intensity exhibits variations in distribution across different population groups

The study revealed that the geometric mean of eggs per gram (EPG) in 515 cases of *C. sinensis* infection stood at 666. Among these cases, the proportions of mild, moderate, and severe infections were 55.25%, 36.58%, and 8.17%, respectively. Specifically, the proportion of moderate and severe infections was higher in males, accounting for 51% of cases, compared to 36.78% in females ($\chi^2 = 10.89$, $P < 0.05$). The farmers exhibited a significantly higher prevalence of moderate to severe infections, with a rate of 46.06%, compared to 10.00% among non-farmers ($\chi^2 = 8.71$, $P < 0.05$). Furthermore, the prevalence of moderate and severe infections increased with age. Specifically, the rate rose from 29.41% in individuals aged 15 to 29 to 45.18% in those aged 30 and older. Meanwhile, there were differences in the infection intensity of *C. sinensis* among groups living at different altitudes. The intensity of moderate to severe infection in low-altitude areas is 50.35% higher than that in mid- and high-altitude areas, 15.48% ($\chi^2 = 34.59$, $P < 0.05$) (Table 1).

Raw fresh fish consumption patterns linked to higher *C. sinensis* infection rates: demographic and behavioral outcomes

Of the 515 infected individuals, 477 (92.62%) reported having consumed raw fresh fish. Notably, no statistically significant difference was observed in the consumption rate of raw fresh fish between males and females. Furthermore, among those who reported consuming raw fresh fish, 470 individuals (98.53%) had been doing so for more than five years. Additionally, 405 individuals (84.91%) reported consuming raw fresh fish at least twice a year (Table 2).

Among the 477 individuals surveyed who consumed raw fresh fish, a significantly higher proportion of those aged > 30 years reported consuming raw fresh fish at least twice a year compared to those aged 15–30 (86.15% vs 46.67%, $\chi^2 = 17.67$, $P < 0.05$). Similarly, individuals aged > 30 years were more likely to report consuming raw fresh fish for ≥ 5 years compared to those aged 15–30 (99.13% vs 80.00%, $\chi^2 = 36.79$, $P < 0.05$). Farmers exhibited significantly higher rates of consuming raw fresh fish at least twice a year (87.09% vs 35.00%, $\chi^2 = 40.57$, $P < 0.05$) and for ≥ 5 years (99.34% vs 80.00%, $\chi^2 = 49.58$, $P < 0.05$) compared to non-farmers. Individuals with a college degree or higher were more likely to consume raw fresh fish at least twice a year (85.71% vs 60.00%, $\chi^2 = 7.50$, $P < 0.05$) and for ≥ 5 years (99.35% vs 73.33%, $\chi^2 = 68.01$, $P < 0.05$) compared to those with other educational backgrounds (Table 2).

Model rankings: GBM leads with optimal performance

To select the optimal machine learning algorithm model, Table 3 not only enumerates the models but also presents their respective rankings based on area under the curve (AUC) and area under the precision-recall curve (AUPRC). Among these models, the most effective model stands out as the GBM, achieving an AUC of 0.993 and an AUPRC of 0.997. The second highest-performing model is the integrated stacked model with an AUC of 0.993 and AUPRC of 0.997. Following this, the independently ranked models are DRF with an AUC of 0.992 and AUPRC of 0.996, DL with an AUC of 0.986 and AUPRC of 0.995, and GLM with an AUC of 0.986 and AUPRC of 0.995, ranked the sixth, twelfth, and thirteenth, respectively. The XRT model is ranked last with an AUC of 0.970 and an AUPRC of 0.989.

Table 1 Demographic and infection intensity characteristics of the screened subjects in Longsheng County

Variables	Objects	No. infection (%)	Geometric mean EPG among infection persons	Infection intensity		
				Mild (%)	Moderate (%)	Severe (%)
Overall	740	515 (69.59)	666	284 (55.25)	188 (36.58)	42 (8.17)
Co-infection						
With nematode	515	25 (4.85)				
Gender						
Male	364	273 (75.00)*	665	132 (48.35)#	108 (39.56)	33 (12.09)
Female	376	242 (64.10)	667	153 (63.22)	80 (33.20)	9 (3.73)
Age (years)						
< 15	10	0 (0)	0	0(0)	0 (0)	0 (0)
15–29	45	17 (37.78)*	681	12 (70.59)	5 (29.41)	0 (0)
30–44	106	71 (66.98)	671	39 (54.93)	22 (30.99)	10 (14.08)
45–59	299	212 (70.90)	653	125 (58.96)	68 (32.22)	19 (9.00)
> 59	280	215 (76.79)	670	109 (50.70)	93 (43.26)	13 (6.05)
Ethnic group						
Han	137	99 (72.26)	653	61 (61.62)	31 (31.63)	7 (7.14)
Zhuang	390	315 (80.77)*	689	157 (49.84)	29 (9.21)	29 (9.21)
Dong	129	77 (59.69)	680	47 (61.04)	24 (31.17)	6 (7.79)
Yao	84	24 (28.57)	659	20 (83.33)	4 (16.67)	0 (0)
Education						
Primary school and below	332	219 (65.96)	664	129 (58.90)	78 (35.78)	12 (5.50)
Junior high school	338	260 (76.92)	671	128 (49.23)	105 (40.38)	27 (10.38)
Senior high school	33	21 (63.64)	677	15 (71.43)	3 (14.29)	3 (14.29)
College degree or above	37	15 (40.54)	671	13 (86.67)	2 (13.33)	0 (0)
Occupation						
Non-farmer	73	20* (27.40)	682	18 (90.00)#	2 (10.00)	0 (0)
Farmer	667	495 (74.21)	664	267 (53.94)	186 (37.58)	42 (8.48)
Raw fresh fish consumption						
No	145	38 (26.21)*	224	26 (68.42)#	11 (28.95)	1 (2.63)
Yes	595	477 (80.17)	666	259 (54.30)	177 (37.11)	41 (8.60)
Frequency of raw fresh fish consumption						
Never	145	38 (26.21)*	384	26 (68.42)#	11 (28.95)	1 (2.3)
1 time/year	190	72 (37.98)	210	67 (93.06)	5 (6.94)	0 (0)
> 1time/year	405	405 (100)	861	192 (47.41)	172 (42.47)	41 (10.12)
Years of raw fresh fish consumption						
Never	145	38 (26.21)*	384	26 (68.42)#	11 (28.95)	1 (2.3)
< 1 year	36	2 (5.56)	24	2 (100)	0 (0)	0 (0)
1–5 years	88	5 (5.68)	162	5 (100)	0 (0)	0 (0)
> 5 years	471	470 (99.79)	715	252 (53.62)	177(37.66)	41 (8.72)
Knowledge of clonorchiasis						
No	307	186 (60.59)*	666	112 (60.22)	66 (35.48)	8 (4.30)
Yes	433	329 (75.98)	664	173 (52.58)	122 (37.08)	34 (10.33)
Elevation						
Low	529	431 (83.69) *	787	214 (49.65)	176 (40.84)	41 (9.51)
Medium	102	53 (10.29)	267	45 (84.91)	7 (13.21)	1 (1.89)
High	109	31 (6.02)	312	26 (83.87) #	5 (16.13)	0 (0)

*A statistically significant disparity in infection rates

#Significant difference in infection severity levels, specifically between mild and high-intensity cases (moderate and severe infections)

Table 2 Consumption patterns of raw fresh fish among various infected individuals in Longsheng County

Variables	Raw fresh fish consumption (%)	Frequency of raw fresh fish consumption (%)		Years of raw fresh fish consumption (%)		
		1 time/year	> 1time/year	< 1 year	1–5 years	> 5 years
Overall(N = 515)	477 (92.62)	72 (15.09)	405 (84.91)	2 (0.42)	5 (1.05)	470 (98.53)
Gender						
Female	211 (44.23)	42 (19.91)	169 (80.09)	2 (0.95)	3 (1.42)	206 (97.63)
Male	266 (55.77)	30 (11.28)	236 (88.72)	0 (0.00)	2 (0.75)	264 (99.25)
Age(years)						
15–29	15 (3.14)	8 (53.33)	7 (46.67)*	2 (13.33)	1 (6.67)	12 (80.00)#
30–44	66 (13.84)	8 (12.12)	58 (87.88)	0 (0.00)	1 (1.52)	65 (98.48)
45–59	202 (42.35)	15 (7.43)	187 (92.57)	0 (0.00)	2 (0.99)	200 (99.01)
> 59	194 (40.67)	41 (21.13)	153 (78.87)	0 (0.00)	1 (0.52)	193 (99.48)
Ethnic group						
Han	92 (19.29)	13 (14.13)	79 (85.87)	1 (1.09)	3 (3.26)	88 (95.65)
Zhuang	288 (60.38)	37 (12.85)	251 (87.15)	0 (0.00)	1 (0.35)	287 (99.65)
Dong	73 (15.30)	9 (12.33)	64 (87.67)	0 (0.00)	1 (1.37)	72 (98.63)
Yao	24 (5.03)	13 (54.17)	11 (45.83)	1 (4.17)	0 (0.00)	23 (95.83)
Occupation						
Non-farmer	20 (4.19)	13 (65.00)	7 (35.00) *	2 (10.00)	2 (10.00)	16 (80.00)#
Farmer	457 (95.81)	59 (12.91)	398 (87.09)	0 (0.00)	3 (0.66)	454 (99.34)
Education						
Primary school and below	197 (41.30)	49 (24.87)	148 (75.13)	0 (0.00)	0 (0.00)	197 (100.00)
Junior high school	245 (51.36)	14 (5.71)	231 (94.29)	0 (0.00)	3 (1.22)	242 (98.78)
Senior high school	20 (4.19)	3 (15.00)	17 (85.00)	0 (0.00)	0 (0.00)	20 (100.00)
College degree or above	15 (3.14)	6 (40.00)	9 (60.00) *	2 (13.33)	2 (13.33)	11 (73.33)#

*A statistically significant disparity in the frequency of raw fresh fish consumption

#A significant difference in years of raw fresh fish consumption

Performance evaluation of the optimal model GBM

Visual representations of the GBM model's performance are provided in Fig. 2a, b, depicting its receiver operating characteristic curve (ROC curve) and precision-recall curve, respectively. These visualizations offer a clear understanding of the model's discriminative ability and its effectiveness in balancing precision and recall. A Pareto chart, illustrating the intricate relationship between the AUC and prediction time for each model, is presented in Fig. 2c. The analysis reveals that GBM exhibits the highest AUC, coupled with a moderate prediction time. In contrast, while the GLM and DL models offer faster processing speed, they fall behind in terms of AUC. Consequently, this study opts for the GBM model, as it strikes an optimal balance between running time and superior AUC performance. Lastly, the learning curve of the optimal model, GBM, is displayed in Fig. 2d. This curve underscores the model's strong fitting capabilities, as evidenced by the converging trends observed in both the training and cross-validation curves.

Raw fresh fish consumption is the principal factor in the prediction model

To further evaluate the main influencing factors in the prediction model, Fig. 3a presents a heat map that visually represents the variable importance. The years of consuming raw fresh fish, frequency of raw fresh fish consumption, elevation, and water distance emerge as the most critical factors, consistently ranking high in terms of importance. In Fig. 3b, the primary variables utilized by the AutoML model, specifically with the GBM algorithm, are highlighted. Years of raw fresh fish consumption tops the list, followed by the frequency of consumption, elevation, water distance, age, ethnic group, education, alcohol, and knowledge of clonorchiasis. Figure 3c illustrates how each variable contributes to the model's prediction of *C. sinensis* infection. Variables closer to a value of 1 indicate a higher probability of infection. Notably, individuals who have consumed raw fresh fish for over 5 years are predominantly clustered on the right side of the zero axis, indicating a strong association between prolonged raw

Table 3 Output of the automated machine learning models that used 12 variables. Model ranks are ordered according to AUPRCs[a]

Rank	Model ID	AUPRC	AUC
1	GBM_4_AutoML_1_20240324_03632	0.997	0.993
2	GBM_3_AutoML_1_20240324_03632	0.997	0.993
3	GBM_2_AutoML_1_20240324_03632	0.997	0.993
4	StackedEnsemble_BestOfFamily_1_AutoML_1_20240324_03632	0.997	0.993
5	GBM_5_AutoML_1_20240324_03632	0.997	0.992
6	DRF_1_AutoML_1_20240324_03632	0.996	0.992
7	GBM_grid_1_AutoML_1_20240324_03632_model_1	0.996	0.991
8	GBM_grid_1_AutoML_1_20240324_03632_model_5	0.996	0.991
9	GBM_grid_1_AutoML_1_20240324_03632_model_3	0.996	0.989
10	GBM_1_AutoML_1_20240324_03632	0.996	0.989
11	GBM_grid_1_AutoML_1_20240324_03632_model_4	0.996	0.988
12	DeepLearning_grid_1_AutoML_1_20240324_03632_model_1	0.995	0.990
13	GLM_1_AutoML_1_20240324_03632	0.995	0.986
14	DeepLearning_1_AutoML_1_20240324_03632	0.994	0.985
15	GBM_grid_1_AutoML_1_20240324_03632_model_2	0.994	0.986
16	StackedEnsemble_AllModels_1_AutoML_1_20240324_03632	0.994	0.989
17	DeepLearning_grid_2_AutoML_1_20240324_03632_model_2	0.994	0.987
18	DeepLearning_grid_2_AutoML_1_20240324_03632_model_1	0.993	0.987
19	DeepLearning_grid_1_AutoML_1_20240324_03632_model_2	0.993	0.982
20	DeepLearning_grid_3_AutoML_1_20240324_03632_model_1	0.993	0.985
21	XRT_1_AutoML_1_20240324_03632	0.989	0.970
22	DeepLearning_grid_3_AutoML_1_20240324_03632_model_2	0.987	0.962

AUPRC area under the precision-recall curve

fresh fish consumption and increased risk of infection. Conversely, those residing in higher altitudes, represented by blue dots, tend to cluster on the left side, suggesting that these individuals are less susceptible to infection. In Fig. 3d, a SHAP explanation for the GBM diagram is provided, further elucidating the impact of specific variable values on *C. sinensis* infection. It reveals that a duration of raw fresh fish consumption exceeding 5 years has the most significant impact on infection risk.

Multiple key variables are significantly associated with *C. sinensis* infection rate

The PDP is presented in Fig. 4, effectively demonstrating the correlation between various variables and the rate of *C. sinensis* infection. In particular, the variables related to the years and frequency of raw fresh fish consumption exhibit a positive association with the infection rate. The altitude and distance from water sources were negatively correlated with infection rates. This PDP provides a comprehensive visualization of how changes in these variables affect the predicted outcome, enabling a more informed interpretation of the model's predictions.

The applicability of the model in multiple scenarios

Lastly, we identified the four most significant variables: years of raw fresh fish consumption, frequency of raw fresh fish consumption, elevation, and water distance. We proceeded to retrain the AutoML model and assess its performance. The results indicate that the GBM model maintains its superior performance, boasting an AUC of 0.992 and an AUPRC of 0.997 (Supplement Table. S1). For a deeper analysis, Supplement Fig. S1 presents the ROC curve, PR curve, Pareto front plot, and learning curve of the model. The best model constructed without the variables related to raw fish consumption, boasts an AUC of 0.774 and an AUPRC of 0.852 (Supplement Table. S2), though the predicted infection risk is significantly lower compared to scenarios where fish consumption is a primary risk factor.

Discussion

Our study conducted a thorough analysis of cross-sectional survey data collected in Longsheng County, Guangxi, China, uncovering a staggering prevalence rate of 69.59% for *C. sinensis* in the region. This figure surpasses the previously

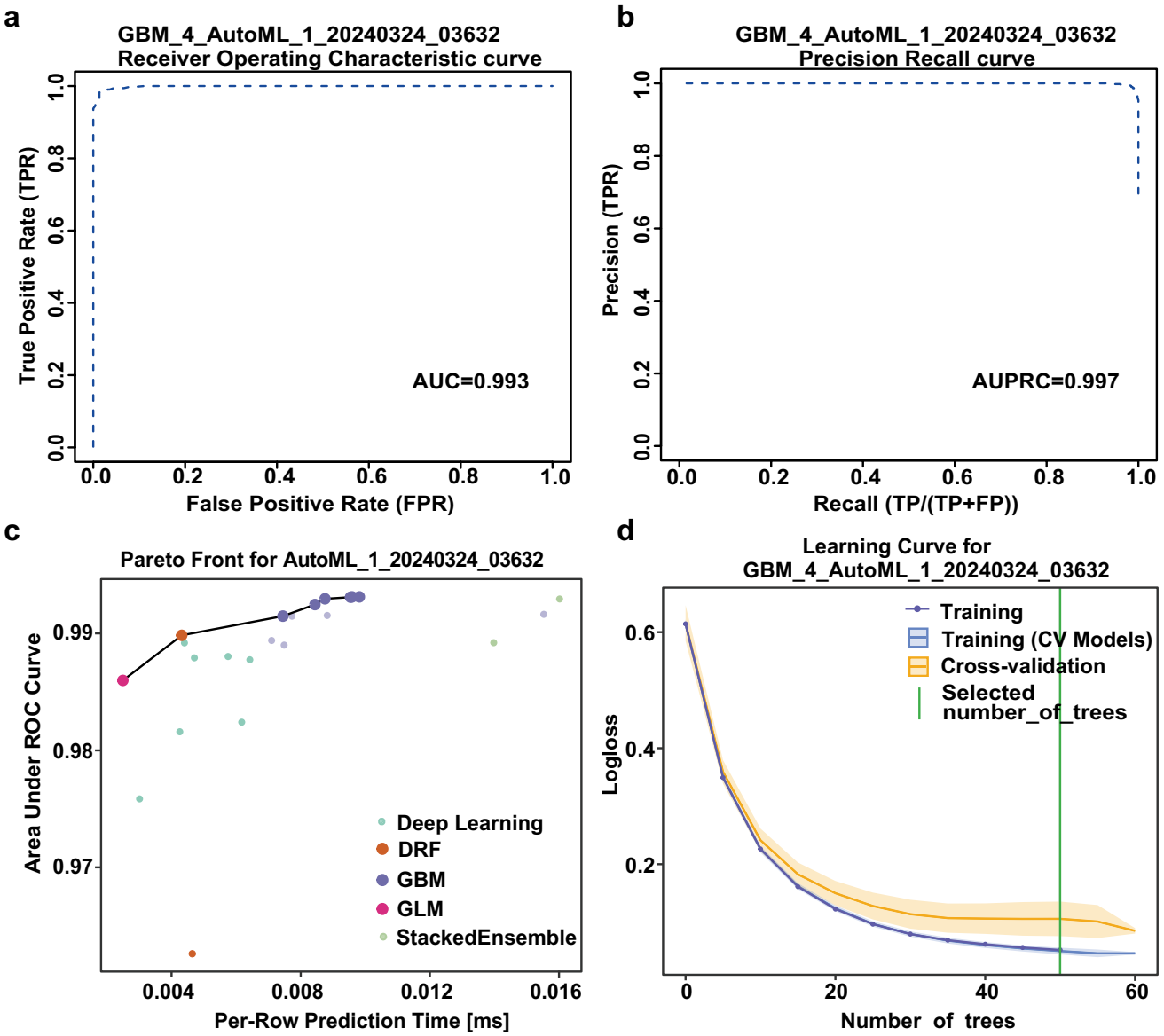

Fig. 2 Model performance evaluation

reported infection rate of 65% for Guangxi in the 2016 national survey (Zhu et al. 2020). According to contemporary parasitological infection classification, Longsheng County is categorized as a severely endemic area for clonorchiasis (Chen 2002). Particularly among individuals aged 30 years and above, particularly those employed in farming, a higher prevalence of infection and a greater likelihood of experiencing moderate to severe symptoms were observed in comparison to other groups. Additionally, men demonstrated a higher prevalence of moderate and severe infections compared to women, echoing findings from studies conducted in both domestic and international regions with a high incidence of *C. sinensis* (Nguyen et al. 2020). This trend could potentially be explained by the longer duration, increased frequency, and larger quantities of raw fresh fish consumed by adult male farmers. And this phenomenon may be attributed to deeply ingrained cultural practices associated with rural lifestyles, where some male participants link the consumption of raw meat to masculinity and strength (Wang et al. 2021).

Despite having some awareness of *C. sinensis*, their community consciousness and dietary habits lead them to continue consuming raw freshwater fish, which they consider an irreplaceable delicacy (Qian et al. 2020). Among the 329 infected individuals, as many as 75.98% continue to eat raw fish, partly due to the widespread misconception that high levels of alcohol, lemon, and vinegar can effectively kill parasites. Additionally, due to the mildness of subjective

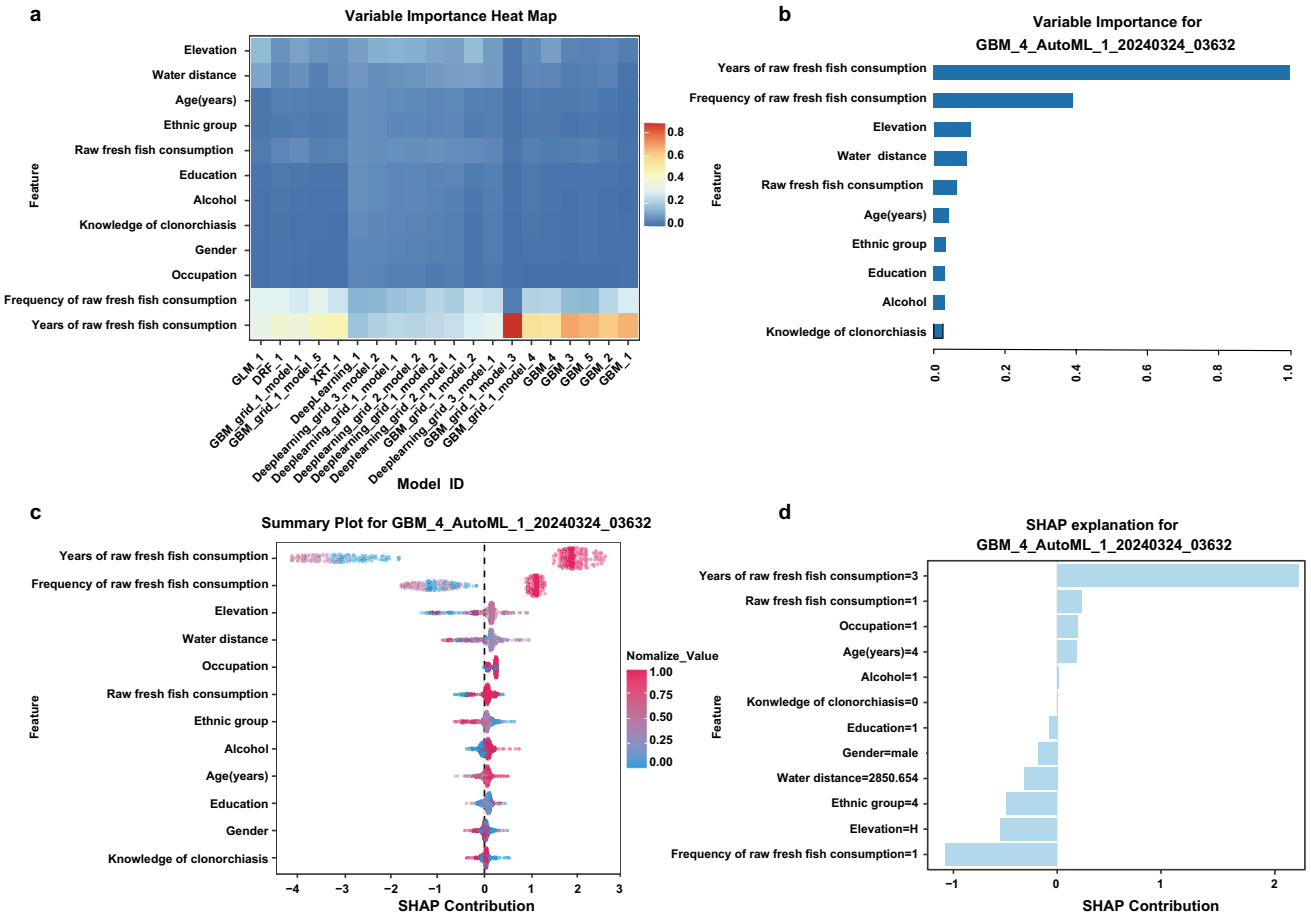

Fig. 3 The interpretability of the model

symptoms, people are unconcerned about the risk of infection and generally believe that *C. sinensis* disease does not lead to serious illness, consistent with previous research (Nguyen et al. 2020; Li et al. 2022). In particular, men tend to believe that taking praziquantel before consumption can prevent infection, making it difficult to stop eating raw freshwater fish even when they attempt to do so (Qian et al. 2022). At the same time, some individuals (38 out of 145, or 26.21%) who did not consume raw fish were still infected with *C. sinensis*. This could be due to respondents not being entirely truthful, or it might be because of other transmission routes, such as water source contamination, contact with infected animals, and kitchen utensils used for preparing raw fish. This phenomenon requires further in-depth research to incorporate environmental monitoring (e.g., testing water sources for *C. sinensis* eggs or metacercariae) and investigate animal reservoirs to provide a more comprehensive understanding of transmission dynamics. Moreover, information, education, and communication (IEC) are often combined with medications to improve the sustainability of control of *C. sinensis* infections (Zhang et al. 2020).

Consuming raw fresh fish is a deeply rooted custom observed during local festivals, with the endorsement of the local government as a means to promote and preserve traditional cultural practices (Qian et al. 2016). This practice is prevalent at social gatherings and dining establishments, where offering raw fresh fish to guests is often regarded as a mark of hospitality. It is noteworthy that adult *C. sinensis* can survive in the human body for numerous years (Tang et al. 2016). Consequently, adult males, who tend to engage more frequently in these customs, exhibit higher exposure levels and parasite loads, leading to increased infection and financial burden (Lin et al. 2024). This correlation between the frequency of consuming raw fresh fish and the intensity of infection is consistent with findings from a study conducted in a community in Shunde District, Guangdong Province, China (Zhang 2016). In Longsheng County, the act of sharing and enjoying traditional raw fresh fish holds significance not only as a way for individuals to establish a sense of cultural heritage but also as a mechanism for fostering social cohesion within the community. When devising interventions, it is imperative to carefully consider the economic benefits derived from the practice of consuming raw

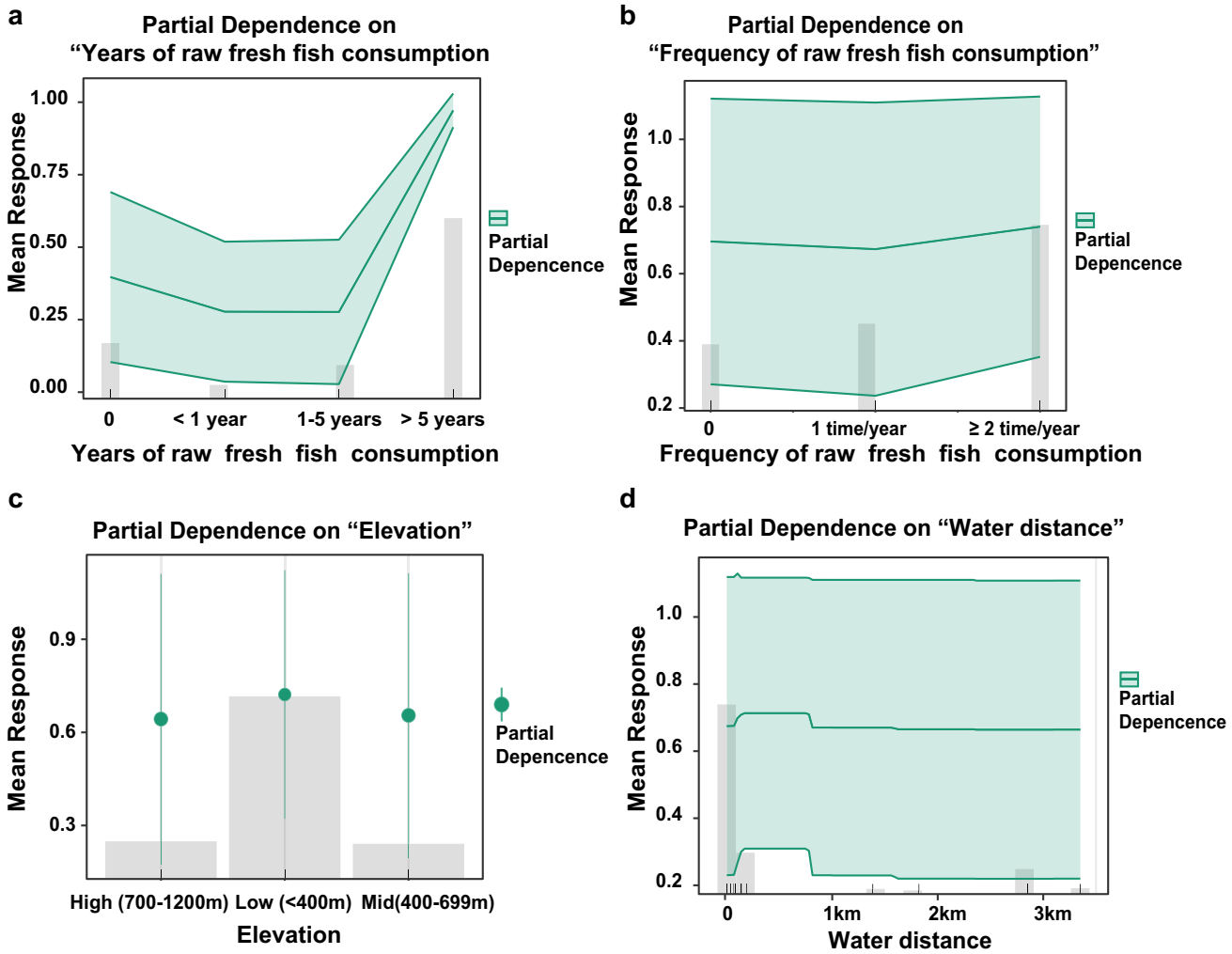

Fig. 4 PDP for the four key variables in the GBM model

fresh fish, including its role in promoting local tourism and supporting traditional industries.

Machine learning and artificial intelligence (AI) have gained significant traction in the distribution of parasitic diseases, owing to their escalating importance. By meticulously analyzing vast epidemiological, environmental, and socioeconomic datasets, these technologies can accurately pinpoint high-risk areas for parasitic diseases and anticipate future infection patterns (Lee et al. 2021). Our research harnessed the AutoML platform to develop and validate a suite of automated machine learning models, in which the GBM (Gradient Boosting Machine) was the most accurate predictor with an AUPRC of 0.997. AutoML demonstrates remarkable efficiency compared to traditional machine learning and statistical analysis, significantly reducing the time required and enhancing accuracy, thereby boosting work productivity (He et al. 2021). Furthermore, the ensemble model integrates various machine learning algorithms and leverages multiple classifiers to forecast target outcomes through a voting system, thereby enhancing overall efficiency (Mienye and Sun 2022). AutoML plays a crucial role in streamlining the labor-intensive stages of the machine learning workflow, such as data preprocessing, model tuning, model selection, result analysis, and model interpretation. By automating these processes, AutoML significantly lowers technical barriers, thereby facilitating the application of machine learning methodologies by public health researchers. We chose the AUPRC as our key metric for assessing model utility, as it directly addresses two crucial performance metrics: positive predictive value and sensitivity. Our objective was to identify patients likely to be infected, enabling us to take proactive measures and treat as many individuals as possible. In contrast, the area under the receiver operating characteristic curve (AUC) evaluates model sensitivity and specificity but fails to consider how infection rates affect performance. The infection rate provides vital context for the

model's effectiveness; without this information, the model loses its relevance.

The importance of explainability in machine learning models cannot be overstated, as it is essential for facilitating inquiry, comprehension, and confidence in AI and machine learning systems. The SHAP chart results indicate that variables like years of raw fresh fish consumption, frequency of raw fresh fish consumption, elevation, and water distance exert a notable influence on *C. sinensis* infection. The PDP further illustrates the quantitative impact of these variables. In the Longsheng region, a direct correlation is evident between the increasing frequency and duration of consuming raw fresh fish and the heightened likelihood of contracting *C. sinensis* infection. Furthermore, as the elevation decreases and water distance diminishes, the risk of *C. sinensis* infection increases. This phenomenon might be attributed to the parasite's preference for vast water bodies and optimal temperatures prevalent in lower-altitude regions, conditions that favor its survival and proliferation. Alternatively, it is conceivable that populations residing at lower elevations and closer to water sources might have greater access to raw fresh fish and consequently consume higher quantities of it, thereby contributing to the observed trend. These findings suggest that efforts should prioritize villagers living at low altitudes near water source areas by raising awareness of prevention and control strategies and mitigating water pollution to reduce the infection rate of *C. sinensis*. This behavior is influenced by social factors and remains consistent across individuals regardless of their occupations, genders, or knowledge levels about *C. sinensis*. Similarly, patterns are observed in other geographic areas as well (Saenna et al. 2017). Convincing people to abandon this culturally entrenched practice poses a significant challenge.

Indeed, consumption of raw freshwater fish remains the primary risk factor for *C. sinensis* infection. However, our inclusion of additional factors—such as water proximity, elevation, and education level—aims to provide a more comprehensive analysis of potential risk contributors in the geographically diverse Longsheng County. Geographical features, including elevation and water accessibility, may indirectly influence infection risk through variations in fish availability, transportation, or dietary practices. Moreover, lower education levels in remote mountainous areas could limit awareness of safe fish consumption, further exacerbating infection rates. To demonstrate the importance of secondary factors, the variables related to raw fish consumption are set to zero; the model relies on factors like elevation, water distance, occupation, ethnic group, alcohol consumption, age, education, gender, and knowledge of *C. sinensis*. The results of these simulations suggest that the best model's prediction (AUC = 0.774, AUPRC = 0.852) is still non-zero, though the predicted infection risk is significantly lower compared to scenarios where fish consumption is a primary risk factor. This analysis helps to clarify the relative importance of raw fish consumption compared to other factors in the context of this study.

Remarkably, a model utilizing only the four most important variables proved effective in predicting *C. sinensis* infection. This outcome underscores the success of our dimensionality reduction process, demonstrating that a model based on four variables can achieve high performance. This finding suggests that not all risk factors are crucial for calculations and making predictions. When initiating infection predictions, researchers should prioritize the utilization of questionnaires, specifically surveys containing these four variables. Dimensionality reduction not only reduces the number of variables to be examined but also minimizes the occurrence of missing values within the dataset. This approach further mitigates the risk of imputation bias, enhancing the reliability of our predictions. Additionally, these simplified models provide researchers with a framework to effectively study diverse groups and replicate our findings, thereby contributing to a deeper understanding of the infection under investigation. This model can also guide disease control personnel in the prevention and control of *C. sinensis* infection. At the same time, SHAP diagrams and PDP diagrams can be used to clarify changes in relevant risk factors and evaluate the comparison of effects before and after prevention and control.

Given the high incidence of *C. sinensis* infection in the Longsheng area, especially in low-altitude and near-water source areas, we need to take comprehensive management measures. Through health education to enhance residents' awareness of the source of infection, transmission, and prevention methods. Strengthen the management of infected people, including regular health check-ups and personalized education. At the same time, environmental sanitation and water source protection should be strengthened to reduce fecal pollution and reduce the risk of disease transmission. In addition, the freshwater fish market is strictly regulated to ensure food safety. Through cross-sectoral cooperation and continuous monitoring, we will optimize prevention and control strategies to protect the health of residents.

This research is subject to certain limitations that must be taken into account. Despite the original objective of encompassing a broad demographic in the survey, the substantial presence of local migrant workers resulted in only 740 individuals completing both the questionnaire and the sample collection. Consequently, the study's outcomes may have been partially influenced by this reduced sample size. Another limitation lies in the diagnostic method used; the Kato-Katz technique exhibits limited sensitivity in low-intensity infections (Hong et al. 2003), which may have resulted in underestimating the true prevalence of *C. sinensis*. To address this, we suggest that future research should combine alternative or complementary diagnostic techniques, such as circulating antigen detection methods,

or molecular techniques such as PCR (Huang et al. 2023), which may offer higher sensitivity for detecting low-intensity infections.

Conclusion

Longsheng County exhibits a notably high overall infection rate and intensity of *C. sinensis*. High rates and intensities of *C. sinensis* infection predominantly affect adults, men, and farmers. Furthermore, we have employed AutoML to create a high-performance prediction model for individual *C. sinensis* infection in Longsheng County. Important variables identified as influencing the risk of *C. sinensis* infection are years of raw fresh fish consumption, frequency of raw fresh fish consumption, elevation, and water distance, which offer crucial information for developing targeted prevention strategies aimed at reducing the risk of infection in this population. We recommend that, in developing intervention measures related to fish consumption in the Longsheng community, efforts should focus on finding a balance between preserving traditional practices and minimizing health risks.

Abbreviations AutoML: Automated machine learning; GBM: Gradient boosting machine; AUC: Area under the curve; AUPRC: Area under the precision-recall curve; EPG: Eggs per gram of feces; DL: Deep learning; GLM: Generalized linear model; DRF: Distributed randomization forest; XRT: Extremely randomized trees; SHAP: SHapley Additive Explanations; PDP: Partial dependence plot

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00436-025-08470-8.

Acknowledgements We are grateful to Guilin Center for Disease Control and Prevention experts for their valuable insights and resources. Additionally, we would like to acknowledge the Longsheng Center for Disease Control and Prevention, for their support and guidance throughout this project. The collaboration and assistance provided by all these individuals and institutions have been instrumental in the successful completion of our research.

Author contribution XWL conceived the study, participated in its design, collected the data, conducted the intricate statistical analysis, and was instrumental in drafting the manuscript. YC participated in its design and conducted the intricate statistical analysis. GYH actively participated in the study design and manuscript preparation. XRS contributed significantly to the study's design and coordination, providing invaluable regional insights. GM was instrumental in coordinating the study and made substantial contributions to both drafting and revising the manuscript. As the corresponding author, XHP supervised the entire project, ensuring its accuracy and integrity, and played a crucial role in finalizing the manuscript. All authors have thoroughly reviewed and approved the final version of the manuscript.

Funding This study was supported by the National Natural Science Foundation of China (82060376) and Natural Science Foundation of Guangxi Zhuang Autonomous Region (No.2024GXNSFAA010039). The funders had no role in study design, data collection, analysis, and interpretation, or in writing the report and the decision to submit the article for publication.

Data availability No datasets were generated or analysed during the current study.

Declarations

Ethics approval This study was approved by the Institutional Review Board of Guilin Medical University (Approval No. GLMC20200301). This study was conducted in accordance with the principles outlined in the Declaration of Helsinki.

Consent to participate All participants provided written informed consent prior to their inclusion in the study. In case of children, the parent or guardian provided written consent and the children consented orally. Participant data were anonymized to ensure confidentiality.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Chen XB (2002) Modern Parasitology. People's Military Medical Publishing House, Beijing

Fürst T, Keiser J, Utzinger J (2012) Global burden of human foodborne trematodiasis: a systematic review and meta-analysis. Lancet Infect Dis 12(3):210–221. https://doi.org/10.1016/S1473-3099(11)70294-8

He X, Zhao KY, Chu XW (2021) AutoML: A survey of the state-of-the-art. Knowl Based Syst 212:106622. https://doi.org/10.1016/j.knosys.2020.106622

Hong ST, Choi MH, Kim CH, Chung BS, Ji Z (2003) The Kato-Katz method is reliable for diagnosis of *Clonorchis sinensis* infection. Diagn Microbiol Infect Dis 47(1):345–347. https://doi.org/10.1016/S0732-8893(03)00113-5

Huang TJ, Li L, Li JH, Li X, Li S, Wang XC, Zhang N, Yu YH, Zhang XC, Zhao ZT, Guo YB, Cao LL, Gong PT (2023) Rapid, sensitive, and visual detection of *Clonorchis sinensis* with an RPA-CRISPR/Cas12a-based dual readout portable platform. Int J Biol Macromol 249:125967. https://doi.org/10.1016/j.ijbiomac.2023.125967

Kim CS, Smith JF, Suwannatrai A, Echaubard P, Wilcox B, Kaewkes S, Sithithaworn P, Sripa B (2017) Role of socio-cultural and economic factors in cyprinid fish distribution networks and consumption in Lawa Lake region, Northeast Thailand: Novel perspectives on *Opisthorchis viverrini* transmission dynamics. Acta Trop 170:85–94. https://doi.org/10.1016/j.actatropica.2017.02.010

Lee SE, Shin HE, Lee MR, Kim YH, Cho SH, Ju JW (2020) Risk Factors of *Clonorchis sinensis* Human Infections in Endemic Areas, Haman-Gun, Republic of Korea: A Case-Control Study. Korean J Parasitol 58(6):647–652. https://doi.org/10.3347/kjp.2020.58.6.647

Lee YW, Choi JW, Shin EH (2021) Machine learning model for diagnostic method prediction in parasitic disease using clinical information. Expert Syst Appl 185:115658. https://doi.org/10.1016/j.eswa.2021.115658

Li Z, Xin H, Qian MB, Sun J, Yang Y, Chen Y, Yu J, Chen Y, Huang Z, Hay SI, Jiang Z, Li SZ (2022) *Clonorchis sinensis* Reinfection rate and reinfection determinants: a prospective cohort study in Hengxian County, Guangxi, China. J Infect Dis 225(3):481–491. https://doi.org/10.1093/infdis/jiab403

Li HM, Zheng JX, Midzi N, Mutsaka-Makuvaza MJ, Lv S, Xia S, Qian YJ, Xiao N, Berguist R, Zhou XN (2024) Schistosomiasis transmission in Zimbabwe: modelling based on machine learning. Infect Dis Model 9(4):1081–1094. https://doi.org/10.1016/j.idm.2024.06.001

Lin DT, Deng ZH, Chen ZB, Jiang KF, Zhang QM, Zhou WJ, Zhang QX, Liu J, Wu ZD, Guo L, Sun X (2024) The disease burden and its distribution characteristics of clonorchiasis in Guangdong Province, Southern China. Parasit Vectors 17(1):353. https://doi.org/10.1186/s13071-024-06425-z

Liu K, Tan J, Xiao L, Pan RT, Yao XY, Shi FY, Li SZ, Li LH (2023) Spatio-temporal disparities of *Clonorchis sinensis* infection in animal hosts in China: a systematic review and meta-analysis. Infect Dis Poverty 12(05):1–31. https://doi.org/10.1186/s40249-023-01146-4

Lun ZR, Gasser RB, Lai DH, Li AX, Zhu XQ, Yu XB, Fang YY (2005) Clonorchiasis: a key foodborne zoonosis in China. Lancet Infect Dis 5(1):31–41. https://doi.org/10.1016/S1473-3099(04)01252-6

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9

Lv GL, Wan XL, Liu J, Ou FQ, Wei HY, Zhang WW, Lin Y (2021) Analysis on clonorchiasis surveillance in Guangxi Zhuang Autonomous Region from 2016 to 2020. J Trop Dis Parasitol 19(3):121

Mienye ID, Sun YX (2022) A survey of ensemble learning: concepts, algorithms, applications, and prospects. IEEE Access 10:99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Milad D, Antaki F, Bernstein A, Touma S, Duval R (2024) Automated machine learning versus expert-designed models in ocular toxoplasmosis: detection and lesion localization using fundus images. Ocul Immunol Inflamm, 1–7. https://doi.org/10.1080/09273948.2024.2319281

Na BK, Pak JH, Hong SJ (2020) *Clonorchis sinensis* and clonorchiasis. Acta Trop 203:105309. https://doi.org/10.1016/j.actatropica.2019.105309

Nguyen TTB, Dermauw V, Dahma H, Bui DT, Le TTH, Phi NTT, Lempereur L, Losson B, Vandenberg O, Do DT, Dorny P (2020) Prevalence and risk factors associated with *Clonorchis sinensis* infections in rural communities in northern Vietnam. PLoS Negl Trop Dis 14(8):e0008483. https://doi.org/10.1371/journal.pntd.0008483

Odongo-Aginya EI, Taylor MG, Sturrock RF, Ackers JP, Doehring E (1995) Field evaluation of an improved Kato-Katz thick smear technique for quantitative determination of helminth eggs in faeces. Trop Med Parasitol 46(4):275–277

Qian MB, Chen YD, Fang YY, Xu LQ, Zhu TJ, Tan T, Zhou CH, Wang GF, Jia TW, Yang GJ, Zhou XN (2011) Disability weight of *Clonorchis sinensis* infection: captured from community study and model simulation. PLoS Negl Trop Dis 5(12):e1377. https://doi.org/10.1371/journal.pntd.0001377

Qian MB, Utzinger J, Keiser J, Zhou XN (2016) Clonorchiasis. Lancet 387(10020):800–810. https://doi.org/10.1016/S0140-6736(15)60313-0

Qian MB, Jiang ZH, Zhou CH, Ge T, Wang X, Zhou XN (2020) Familial assimilation in transmission of raw-freshwater fish-eating practice leading to clonorchiasis. PLoS Negl Trop Dis 14(4):e0008263. https://doi.org/10.1371/journal.pntd.0008263

Qian MB, Zhou CH, Jiang ZH, Yang YC, Lu MF, Wei K, Wei SL, Chen Y, Li HM, Zhou XN (2022) Epidemiology and determinants of *Clonorchis sinensis* infection: a community-based study in southeastern China. Acta Trop 233:106545. https://doi.org/10.1016/j.actatropica.2022.106545

Qian MB, Keiser J, Utzinger J, Zhou XN (2024) Clonorchiasis and opisthorchiasis: epidemiology, transmission, clinical features, morbidity, diagnosis, treatment, and control. Clin Microbiol Rev 37:e00009-23. https://doi.org/10.1128/cmr.00009-23

Roessler AS, Oehm AW, Knubben-Schweizer G, Groll A (2022) A machine learning approach for modelling the occurrence of *Galba truncatula* as the major intermediate host for *Fasciola hepatica* in Switzerland. Prev Vet Med 200:105569. https://doi.org/10.1016/j.prevetmed.2022.105569

Rong GX, Wu JR, Yang Z, Huang YM (2011) Epidemiology survey on clonorchiasis sinensis in Longsheng Minority Autonomous Counties of Guangxi. Chin J Public Health Manag 27(4):392–393. https://doi.org/10.19568/j.cnki.23-1318.2011.04.032

Saenna P, Hurst C, Echaubard P, Wilcox BA, Sripa B (2017) Fish sharing as a risk factor for *Opisthorchis viverrini* infection: evidence from two villages in north-eastern Thailand. Infect Dis Poverty 6(1):66. https://doi.org/10.1186/s40249-017-0281-7

Sripa B, Suwannatrai AT, Sayasone S, Do DT, Khieu V, Yang Y (2021) Current status of human liver fluke infections in the Greater Mekong Subregion. Acta Trop 224:106133. https://doi.org/10.1016/j.actatropica.2021.106133

Sun JL, Xin HH, Qian MB, Duan KX, Chen YD, Li SZ, Li W, Huang SY, Gan XQ, Yang YC, Li ZJ (2020) High endemicity of *Clonorchis sinensis* infection in Binyang County, southern China. PLoS Negl Trop Dis 14(8):e0008540. https://doi.org/10.1371/journal.pntd.0008540

Tang ZL, Huang Y, Yu XB (2016) Current status and perspectives of *Clonorchis sinensis* and clonorchiasis: epidemiology, pathogenesis, omics, prevention and control. Infect Dis Poverty 5(1):71. https://doi.org/10.1186/s40249-016-0166-1

Vinh HQ, Phimpraphai W, Tangkawattana S, Smith JF, Kaewkes S, Dung DT, Duong TT, Sripa B (2017) Risk factors for *Clonorchis sinensis* infection transmission in humans in northern Vietnam: a descriptive and social network analysis study. Parasitol Int 66(2):74–82. https://doi.org/10.1016/j.parint.2016.11.018

Wan XL, Zhang WW, Jiang ZH, Lv GL, Ou FQ, Wei HY, Lin Y, Tang WQ, Wei SJ, Huang KL (2019) Investigation on the status of human important parasitic disease in Guangxi in 2015. China Trop Med 19(1):19–30. https://doi.org/10.13604/j.cnki.46-1064/r.2019.01.05

Wang YC, Grundy-Warr C, Namsanor J, Kenney-Lazar M, Tang CJY, Goh LYW, Chong YC, Sithithaworn P, Ngonkum S, Khuntikeo N (2021) Masculinity and misinformation: Social dynamics of liver fluke infection risk in Thailand. Parasitol Int 84:102382. https://doi.org/10.1016/j.parint.2021.102382

Wang YC, Namsanor J, Law A, Sithithaworn P (2023) A socio-ecological framework for examining foodborne parasitic infection risk. Acta Trop 244:106957. https://doi.org/10.1016/j.actatropica.2023.106957

Xin HL, Yang YC, Jiang ZH, Qian MB, Chen YD, Li SZ, Cowling BJ, Sun JL, Li ZJ (2021) An investigation of Human Clonorchiasis prevalence in an Endemic County in Guangxi Zhuang Autonomous Region, China, 2016. Food Waterborne Parasitol 22:e00109. https://doi.org/10.1016/j.fawpar.2020.e00109

Xu M, Jiang YY, Yin JH, Cao SK, Shen YJ, Cao JP (2021) Risk factors for *clonorchis sinensis* infection in residents of Binyang, Guangxi: a cross-sectional and logistic analysis study. Front Public Health 9:588325. https://doi.org/10.3389/fpubh.2021.588325

Xu N, Zhang Y, Du CH, Song J, Huang JH, Gong YF, Jiang HL, Tong YX, Yin JF, Wang JM, Jiang F, Chen Y, Jiang QW, Dong Y, Zhou YB (2023) Prediction of *Oncomelania hupensis* distribution in association with climate change using machine learning models. ParasitVectors 16(1):377. https://doi.org/10.1186/s13071-023-05952-5

Yang JX, Tan CY, Wei HY, Jiang ZH, Yang XA (2017) Wang XL (2019) Status of *Clonorchis sinensis* in the minority area of northern Guangxi. China Trop Med 19(6):571–573. https://doi.org/10.13604/j.cnki.46-1064/r.2019.06.17

Zhan SY (2012) Epidemiology. People's Medical Publishing House, Beijing, pp 64–65

Zhang W (2016) Infection status of liver fluke disease and its epidemiological factors among population in Shunde District of Foshan City in 2015. Occup Health 32:1662–1664. https://doi.org/10.13329/j.cnki.zyyjk.2016.0477

Zhang LJ, Zhu HQ, Wang Q, Lu S, Xu J, Li SZ (2020) Assessment of schistosomiasis transmission risk along the Yangtze River basin after the flood disaster in 2020. Chin J Schistosomiasis Control 32(5):464–468. https://doi.org/10.16250/j.32.1374.2020242

Zhao LQ, Lin DR, Lin HT (2024) Automated machine learning for diabetic retinopathy progression. JAMA Ophthalmol 142(3):178–179. https://doi.org/10.1001/jamaophthalmol.2023.6778

Zheng JX, Zhu HH, Xia S, Qian MB, Nguyen HM, Sripa B, Sayasone S, Khieu V, Bergquist R, Zhou XN (2024) Natural variables separate the endemic areas of *Clonorchis sinensis* and *Opisthorchis viverrini* along a continuous, straight zone in Southeast Asia. Infect Dis Poverty 13(1):24. https://doi.org/10.1186/s40249-024-01191-7

Zhu TJ, Chen YD, Qian MB, Zhu HH, Huang JL, Zhou CH, Zhou XN (2020) Surveillance of clonorchiasis in China in 2016. Acta Trop 203:105320. https://doi.org/10.1016/j.actatropica.2019.105320