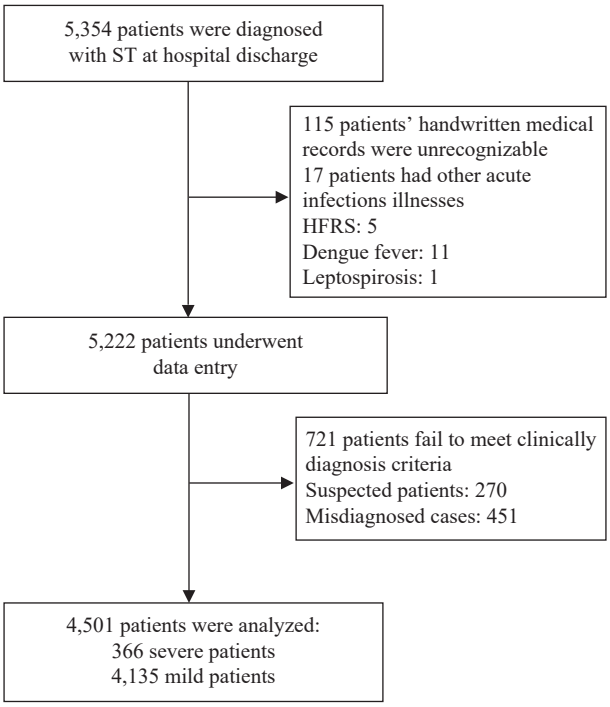
SUPPLEMENTARY METHODS

Steps of Multivariate Logistic Regression Analysis

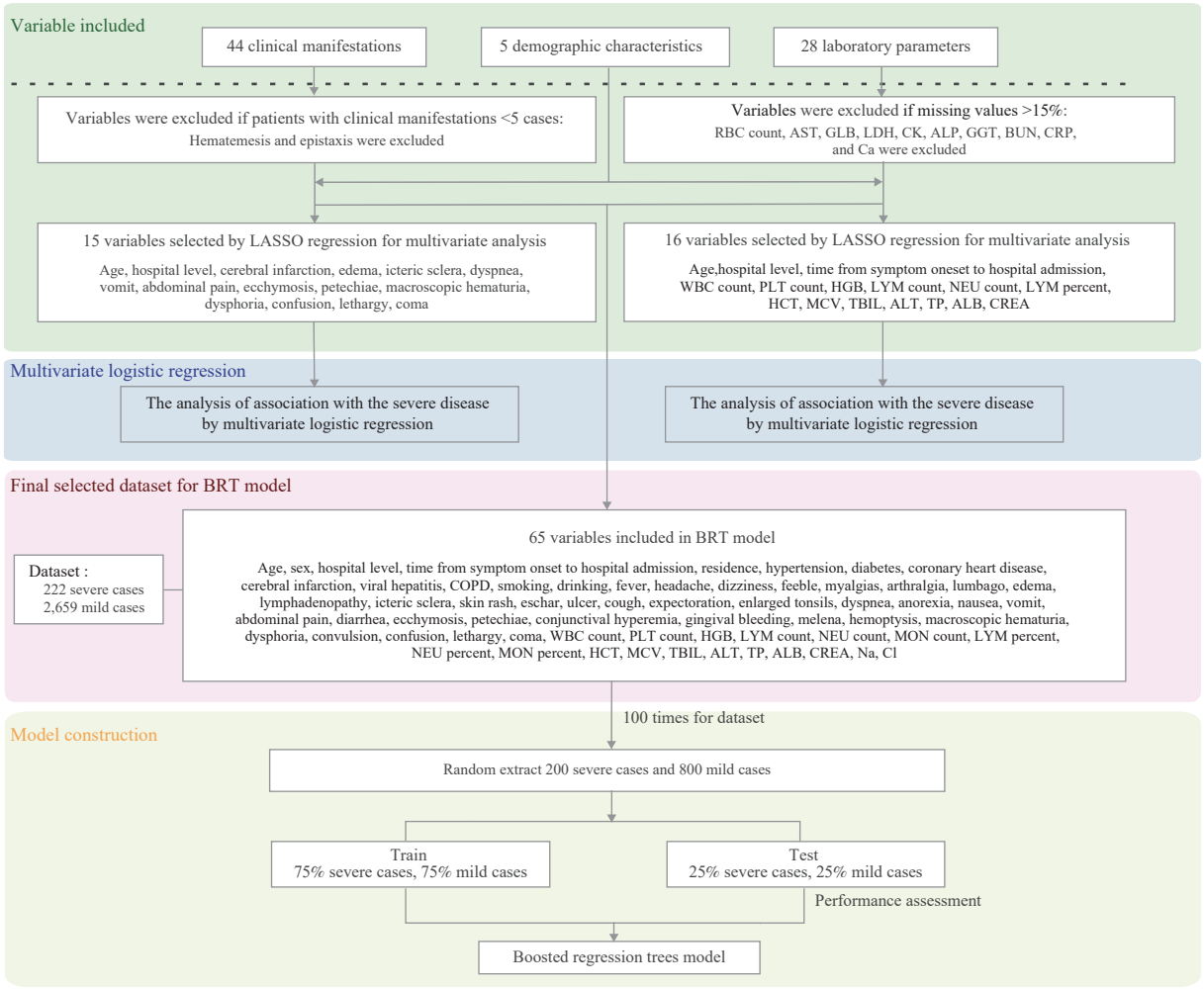
At the first step, to determine the risk factors of clinical manifestations in severe scrub typhus (ST), the presence of clinical manifestations on or before hospital admission was used for statistical analysis. Excluding variables of clinical manifestations was considered if the number of patients with a single clinical manifestation were less than five cases, and hematemesis and epistaxis were excluded. The remaining clinical manifestations, together with 5 demographic characteristics were included in Least Absolute Shrinkage and Selection Operator (LASSO) regression to select the strongest variables. Variables identified by LASSO regression analysis were entered into multivariate logistic regression analysis (Supplementary Figure S2). Meanwhile, to attain the risk factors of laboratory indicators in severe ST, the laboratory indicators result within 24 hours of hospital admission were used for statistical analysis. After excluding the laboratory indicators if the missing value was more than 15%, the laboratory indicators were also examined for their association with development of severe disease in multivariate logistic regression analysis after variable selection of LASSO regression analysis (Supplementary Figure S2). The R library "glmnet" was used for the LASSO regression. Odds ratio (OR) and the 95% confidence interval (95% CI) were estimated using maximum likelihood methods.

Variable Selection for Model Construction of the Boosted Regression Trees (BRT) Modeling

To attain an early prediction of severe ST, the clinical manifestations on or before hospital admission and the laboratory indicators result within 24 hours of hospital admission were used for model analysis. Firstly, for 44 variables of clinical manifestations, we excluded those less than five cases such as hematemesis and epistaxis. Meanwhile, for 28 laboratory parameters, we excluded those with missing values more than 15% of all patients such as RBC count, AST, GLB, LDH, CK, ALP, GGT, BUN, CRP, and Ca (Supplementary Figure S2). The remaining clinical manifestations and laboratory parameters, together with 5 demographic characteristics were included in the BRT models.



SUPPLEMENTARY FIGURE S1. The flowchart of recruited patients with scrub typhus in Guangzhou, China, 2012–2018. Abbreviations: ST=scrub typhus; HFRS=hemorrhagic fever with renal syndrome.



SUPPLEMENTARY FIGURE S2. Variable selection and model construction of the logistic regression and boosted regression trees modeling.

Notes: Demographic characteristics including age, sex, hospital level, time from symptom onset to hospital admission, and residence. Clinical manifestations including hypertension, diabetes, coronary heart disease, cerebral infarction, viral hepatitis, COPD, smoking, drinking, fever, headache, dizziness, weakness, myalgias, arthralgia, lumbago, edema, lymphadenopathy, icteric sclera, skin rash, eschar, ulcer, cough, expectoration, enlarged tonsils, dyspnea, anorexia, nausea, vomit, abdominal pain, diarrhea, ecchymosis, petechiae, conjunctival hyperemia, gingival bleeding, melena, hemoptysis, hematemesis, epistaxis, macroscopic hematuria, dysphoria, convulsion, confusion, lethargy, and coma. Laboratory parameters including WBC count, RBC count, PLT count, HGB, LYM count, NEU count, MON count, LYM percent, NEU percent, MON percent, HCT, MCV, TBIL, AST, ALT, TP, ALB, GLB, LDH, CK, ALP, GGT, CREA, BUN, CRP, Na, Cl, and Ca. Residence, rural or urban were divided according to the county address of patients, in where Huadu, Conghua, Zengcheng, Nansha, and Panyu districts were considered as rural areas, and Tianhe, Yuexiu, Liwan, Haizhu, Baiyun, and Huangpu districts were considered as urban areas (*5*).

Abbreviations: RBC=red blood cells; AST=aspartate aminotransferase; GLB=globulin; LDH=lactate dehydrogenase; CK=creatine kinase; ALP=alkaline phosphatase; GGT=γ-glutamyl transpeptidase; BUN=blood urea nitrogen; CRP=C-creative protein; Ca=calcium; LASSO=Least Absolute Shrinkage and Selection Operator; WBC=white blood cells; PLT=platelet; HGB=hemoglobin; LYM=lymphocyte; NEU=neutrophil; HCT=hematocrit; MCV=mean corpuscular volume; TBIL=total bilirubin; ALT=alanine aminotransferase; TP=total protein; ALB=albumin; CREA=creatinine; BRT=boosted regression trees; COPD=chronic obstructive pulmonary disease; MON=monocyte; Na=sodium; Cl=chlorine.

Steps of BRT Modeling

Machine learning techniques are increasingly used in developing models for vector predictions with multiple advantages, like considering nonlinearity and interactions and handling different types of predictor variables (*1–2*).

The BRT modeling was carried out in steps. First, considering the low proportion of severe outcomes, we set a 1-to-4 case-control ratio to adjust every bootstrap data. Severe cases and 4-fold num of mild cases were randomly

SUPPLEMENTARY TABLE S1. Demographic and clinical characteristics associated with severe disease for patients with scrub typhus by multivariate logistic regression.

Variables	Total (n=4,501)	Severe (n=366)	Mild (n=4,135)	Adjusted OR (95% CI)	*P* value
Sex, female	2,450 (54.4)	198 (54.1)	2,252 (54.5)		
Age, years, median (IQR)	57 (46–65)	60 (51–68)	56 (46–64)	1.01 (1.01–1.02)	0.001
Time from symptom onset to hospital admission, days, median (IQR)	7 (5–9)	8 (6–10)	7 (4–9)	2.76 (1.91–4.00)	<0.001
Length of hospital stay, days, median (IQR)	7 (6–10)	12 (8–17)	7 (5–9)		
Comorbidities					
Hypertension	718 (16.0)	77 (21.0)	641 (15.5)		
Diabetes	372 (8.3)	38 (10.4)	334 (8.1)		
Coronary heart disease	112 (2.5)	8 (2.2)	104 (2.5)		
Cerebral infarction	136 (3.0)	32 (8.7)	104 (2.5)	2.33 (1.37–3.98)	0.002
Viral hepatitis	178 (4.0)	12 (3.3)	166 (4.0)		
COPD	46 (1.0)	6 (1.6)	40 (1.0)		
Lifestyles					
Smoking	731 (16.2)	64 (17.5)	667 (16.1)		
Drinking	384 (8.5)	37 (10.1)	347 (8.4)		
Non-specific					
Fever	4,421 (98.2)	356 (97.3)	4,065 (98.3)		
Headache	2,302 (51.1)	174 (47.5)	2,128 (51.5)		
Dizziness	1,168 (25.9)	90 (24.6)	1,078 (26.1)		
Weakness	1,969 (43.7)	155 (42.3)	1,814 (43.9)		
Myalgias	1,051 (23.4)	82 (22.4)	969 (23.4)		
Arthralgia	136 (3.0)	5 (1.4)	131 (3.2)		
Lumbago	140 (3.1)	9 (2.5)	131 (3.2)		
Edema	178 (4.0)	64 (17.5)	114 (2.8)	4.85 (3.25–7.24)	<0.001
Lymphadenopathy	1,088 (24.2)	92 (25.1)	996 (24.1)		
Icteric sclera	79 (1.8)	33 (9.0)	46 (1.1)	4.67 (2.61–8.33)	<0.001
Skin					
Skin rash	653 (14.5)	61 (16.7)	592 (14.3)		
Eschar/Ulcer	3,921 (87.1)	302 (82.5)	3,619 (87.5)		
Eschar	3,335 (74.1)	260 (71.0)	3,075 (74.4)		
Ulcer	784 (17.4)	60 (16.4)	724 (17.5)		
Respiratory					
Cough	1,858 (41.3)	169 (46.2)	1,689 (40.8)		
Expectoration	1,157 (25.7)	116 (31.7)	1,041 (25.2)		
Enlarged tonsils	426 (9.5)	30 (8.2)	396 (9.6)		
Dyspnea	213 (4.7)	126 (34.4)	87 (2.1)	13.95 (9.94–19.58)	<0.001
Gastrointestinal					
Anorexia	3,148 (69.9)	270 (73.8)	2,878 (69.6)		
Nausea	791 (17.6)	69 (18.9)	722 (17.5)		
Vomit	484 (10.8)	63 (17.2)	421 (10.2)	1.45 (1.02–2.04)	0.036
Abdominal pain	361 (8.0)	64 (17.5)	297 (7.2)	1.90 (1.32–2.73)	0.001
Diarrhea	193 (4.3)	35 (9.6)	158 (3.8)		

TABLE S1. (Continued)

Variables	Total (n=4,501)	Severe (n=366)	Mild (n=4,135)	Adjusted OR (95% CI)	*P* value
Hemorrhagic manifestation*	241 (5.4)	47 (12.8)	194 (4.7)		
Ecchymosis	47 (1.0)	14 (3.8)	33 (0.8)		
Petechiae	22 (0.5)	7 (1.9)	15 (0.4)	3.45 (1.05–11.38)	0.042
Conjunctival hyperemia	108 (2.4)	20 (5.5)	88 (2.1)		
Gingival bleeding	9 (0.2)	2 (0.5)	7 (0.2)		
Melena	59 (1.3)	10 (2.7)	49 (1.2)		
Hemoptysis	10 (0.2)	3 (0.8)	7 (0.2)		
Hematemesis	3 (0.1)	1 (0.3)	2 (<0.1)		
Epistaxis	4 (0.1)	0 (0)	4 (0.1)		
Macroscopic hematuria	8 (0.2)	3 (0.8)	5 (0.1)	5.32 (1.03–27.39)	0.045
Neurological manifestation†	124 (2.8)	70 (19.3)	54 (1.3)		
Dysphoria	29 (0.6)	19 (5.2)	10 (0.2)	6.76 (2.40–19.10)	<0.001
Convulsion	27 (0.6)	15 (4.1)	12 (0.3)		
Confusion	50 (1.1)	34 (9.3)	16 (0.4)	7.18 (3.29–15.67)	<0.001
Lethargy	31 (0.7)	18 (4.9)	13 (0.3)	5.80 (2.19–15.41)	<0.001
Coma	30 (0.7)	18 (4.9)	12 (0.3)	2.90 (1.08–7.79)	0.035
Abnormal image results					
Pericardial effusion	127 (2.8)	37 (10.1)	90 (2.2)		
Pelvic effusion	42 (0.9)	14 (3.8)	28 (0.7)		
Pleural effusion	558 (12.4)	158 (43.2)	400 (9.7)		
Chest radiographic abnormality	1,062 (23.6)	167 (45.6)	895 (21.6)		
Ascites	50 (1.1)	26 (7.1)	24 (0.6)		
Splenomegaly	554 (12.3)	74 (20.2)	480 (11.6)		
Hepatomegaly	228 (5.1)	44 (12.0)	184 (4.4)		

Note: Data are n (%) until otherwise indicated. All clinical symptoms (non-specific, skin, respiratory, gastrointestinal, hemorrhagic, neurological) were reported before or at hospital admission, all abnormal image results were reported after hospital admission.
Abbreviations: COPD=chronic obstructive pulmonary disease. OR=odds ratio. CI=confidence interval; IQR=interquartile-range.
* Hemorrhagic manifestation, patients with one or more hemorrhagic symptoms.
† Neurological manifestation, patients with one or more neurological symptoms.

selected without replacement. Second, for both selected severe and mild cases, 75% of samples of each were used for training the models while the rest 25% were used for model testing. Third, a BRT model was built using the training dataset and validated using the test dataset.

To increase the predictive robustness of model predictions and quantify model uncertainty, we fitted an ensemble of 100 BRT models to separate bootstraps of the data. We adopted a tree complexity of 5, a learning rate of 0.005, and a bag fraction of 75% to identify the optimal trees for each bootstrap data. Herein, we modeled the severe complications as a binary classification. Each of the 100 models predicts severe risk on a continuous scale from 0 to 1, with a final prediction being generated by calculating the mean prediction across all models. To further determine the main contributing factor, we performed BRT model again after removing the factors that showed the relative contributions below 2% (*3*). The receiver-operating characteristic (ROC) curves were produced for each model based on the average of an ensemble of 100 BRT models, and the area under the receiver operating characteristic curve (AUC) was calculated to evaluate the predictive power of the models. BRT model was performed using R software (version 3.6.2; R Foundation; Vienna, Austria) with the packages of "gbm," "dismo," and "ROCR" (*4*).

SUPPLEMENTARY TABLE S2. Laboratory findings on hospital admission compared between severe patients and mild patients.

Variables	Total (n=4,501)	Severe (n=366)	Mild (n=4,135)	*P* value
Hematological indicators				
WBC count (×10^9/L) [4–10]	7.18 (5.23–9.60)	8.99 (6.50–12.92)	7.06 (5.19–9.39)	<0.001
RBC count (×10^{12}/L) [3.8–5.8]	4.23 (3.87–4.64)	3.95 (3.44–4.31)	4.26 (3.89–4.66)	<0.001
PLT count (×10^9/L) [100–300]	121.00 (81.00–172.00)	60.00 (40.00–95.50)	125.00 (86.05–178.00)	<0.001
HGB (g/L) [120–165]	121.00 (110.00–132.00)	110.90 (97.00–124.00)	121.00 (111.00–133.00)	<0.001
LYM count (×10^9/L) [1.1–3.2]	1.42 (0.78–2.57)	1.40 (0.73–2.66)	1.42 (0.79–2.56)	0.672
NEU count (×10^9/L) [1.8–6.3]	4.67 (3.23–6.55)	6.50 (4.47–9.7)	4.55 (3.17–6.36)	<0.001
MON count (×10^9/L) [0.1–0.6]	0.44 (0.28–0.69)	0.42 (0.26–0.74)	0.45 (0.28–0.69)	0.719
LYM percent (%) [20–40]	20.50 (12.90–32.30)	15.00 (8.60–26.90)	21.00 (13.30–32.80)	<0.001
NEU percent (%) [50–70]↑	71.40 (57.50–80.40)	78.40 (63.50–86.30)	71.00 (57.00–79.80)	<0.001
MON percent (%) [3–10]	6.40 (4.40–8.80)	4.90 (3.30–7.50)	6.50 (4.50–8.94)	<0.001
HCT (%) [40–50]↓	35.90 (32.50–39.20)	32.80 (28.60–35.70)	36.07 (32.80–39.40)	<0.001
MCV (FL) [82–100]	86.30 (81.70–90.01)	84.65 (79.50–89.00)	86.50 (81.90–90.20)	<0.001
Biochemical indicators				
TBIL (μmol/L) [5.1–17.1]	11.50 (8.20–16.50)	18.50 (12.38–48.35)	11.00 (8.00–15.80)	<0.001
AST (U/L) [0–40]↑	76.00 (49.00–122.00)	136.00 (88.10–218.40)	73.00 (48.00–115.00)	<0.001
ALT (U/L) [0–40]↑	67.00 (42.00–107.40)	79.50 (54.00–118.00)	65.65 (41.00–106.00)	<0.001
TP (g/L) [65–85]↓	61.20 (56.60–66.10)	54.35 (49.20–60.60)	61.80 (57.30–66.40)	<0.001
ALB (g/L) [35–55]↓	33.00 (28.90–36.70)	25.90 (22.00–29.70)	33.40 (29.60–37.00)	<0.001
GLB (g/L) [20–40]	28.20 (25.10–31.70)	29.00 (25.30–32.60)	28.11 (25.10–31.60)	0.176
LDH (U/L) [109–245]↑	459.00 (342.00–693.00)	677.00 (464.00–1,147.00)	446.50 (336.00–658.00)	<0.001
CK (U/L) [25–200]	91.00 (50.00–173.00)	110.00 (59.00–286.00)	90.00 (50.00–167.00)	<0.001
ALP (U/L) [40–150]	94.00 (67.03–143.00)	127.00 (76.00–220.50)	92.00 (67.00–138.00)	<0.001
GGT (U/L) [7–50]↑	70.00 (33.00–146.00)	98.00 (44.00–189.00)	67.00 (32.00–142.00)	<0.001
CREA (μmol/L) [53–106]	78.08 (64.00–98.00)	102.95 (72.00–193.50)	77.85 (63.10–95.10)	<0.001
BUN (mmol/L) [3.2–7.1]	4.34 (3.30–6.04)	9.00 (5.97–16.60)	4.27 (3.24–5.70)	<0.001
CRP (mg/L) [0.07–8.20]↑	45.00 (14.18–86.36)	84.05 (39.00–140.70)	42.00 (13.31–81.10)	<0.001
Na (mmol/L) [137–147]↓	135.70 (132.00–138.60)	133.00 (129.80–137.00)	136.00 (132.40–138.70)	<0.001
Cl (mmol/L) [96–108]	99.70 (96.40–102.90)	100.00 (96.00–103.90)	99.70 (96.40–102.80)	0.272
Ca (mmol/L) [2.10–2.55]↓	2.07 (1.96–2.18)	1.90 (1.79–2.03)	2.08 (1.98–2.19)	<0.001

Note: All measurements within 24 hours of hospital admission were presented as median (interquartile-range).
P value was compared between severe group and mild group by Mann-Whitney U test.
Abbreviations: WBC=white blood cells; RBC=red blood cells; PLT=platelet; HGB=hemoglobin; LYM=lymphocyte; NEU=neutrophil; MON=monocyte; HCT=hematocrit; MCV=mean corpuscular volume; TBIL=total bilirubin; AST=aspartate aminotransferase; ALT=alanine aminotransferase; TP=total protein; ALB=albumin; GLB=globulin; LDH=lactate dehydrogenase; CK=creatine kinase; ALP=alkaline phosphatase; GGT=γ-glutamyl transpeptidase; CREA=creatinine; BUN=blood urea nitrogen; CRP=C-creative protein; Na=sodium; Cl=chlorine; Ca=calcium.
↑Variable, with the median value in total population above the normal range.
↓Variable, with the median value in total population below the normal range.

SUPPLEMENTARY TABLE S3. Association between laboratory measurements and severe scrub typhus by multivariate logistic regression analysis.

Variables	Adjusted OR (95% CI)	*P* value
Hospital level	2.26 (1.50–3.43)	<0.001
Time from symptom onset to hospital admission, days	1.06 (1.01–1.10)	0.014
PLT count <100 (×10^9/L) [100–300]	4.80 (3.39–6.82)	<0.001
HGB <120 (g/L) [120–165]	2.19 (1.59–3.01)	<0.001
NEU count >6.3 (×10^9/L) [1.8–6.3]	2.03 (1.47–2.79)	<0.001
TBIL >17.1 (μmol/L) [5.1–17.1]	2.62 (1.91–3.60)	<0.001
ALB <35 (g/L) [35–55]	3.78 (1.86–7.65)	<0.001
CREA >106 (μmol/L) [53–106]	2.91 (2.12–4.00)	<0.001

Abbreviations: OR=odds ratio; CI=confidence interval; PLT=platelet; HGB=hemoglobin; NEU=neutrophil; TBIL=total bilirubin; ALB=albumin; CREA=creatinine.

SUPPLEMENTARY TABLE S4. Relative contribution of predictors estimated for the occurrence of severe complications of scrub typhus by the boosted regression trees model.

Variables	Relative contribution, % (95% CI)
ALB (g/L)	24.77 (23.70–25.85)
Dyspnea	16.63 (15.73–17.53)
PLT count (×10^9/L)	12.95 (12.16–13.74)
CREA (μmol/L)	11.34 (10.58–12.11)
TBIL (μmol/L)	10.55 (10.08–11.02)
NEU count (×10^9/L)	5.71 (5.40–6.02)
TP (g/L)	5.59 (5.20–5.98)
NEU percent (%)	5.40 (5.06–5.75)
HGB (g/L)	4.25 (3.97–4.54)
Age	2.79 (2.59–2.98)

Abbreviations: ALB=albumin; PLT=platelet; CREA=creatinine; TBIL=total bilirubin; NEU=neutrophil; TP=total protein; HGB=hemoglobin; CI=confidence interval.

REFERENCES

1. Sinka ME, Rubio-Palis Y, Manguin S, Patil AP, Temperley WH, Gething PW, et al. The dominant *Anopheles* vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic précis. Parasit Vectors 2010;3:72. http://dx.doi.org/10.1186/1756-3305-3-72.
2. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol 2008;77(4):802 – 13. http://dx.doi.org/10.1111/j.1365-2656.2008.01390.x.
3. Zhao GP, Wang YX, Fan ZW, Ji Y, Liu MJ, Zhang WH, et al. Mapping ticks and tick-borne pathogens in China. Nat Commun 2021;12(1):1075. http://dx.doi.org/10.1038/s41467-021-21375-1.
4. Araújo MB, New M. Ensemble forecasting of species distributions. Trends Ecol Evol 2007;22(1):42 – 7. http://dx.doi.org/10.1016/j.tree.2006.09.010.
5. Wei YH, Huang Y, Luo L, Xiao XC, Liu L, Yang ZC. Rapid increase of scrub typhus: an epidemiology and spatial-temporal cluster analysis in Guangzhou City, Southern China, 2006-2012. PLoS One 2014;9(7):e101976. http://dx.doi.org/10.1371/journal.pone.0101976.