The role of genetic polymorphism in predictive model of ATDH

Jingwei Zhang
Chengdu Second People's Hospital
Wei Zhou
University of Electronic Science and Technology of China
Zhenyang Wang
University of Electronic Science and Technology of China
Ping Luo
Sichuan Academy of Medical Sciences, University of Electronic Science and Technology of China
Wang Xiang
Sichuan Academy of Medical Sciences, University of Electronic Science and Technology of China
Xiao Liang
Sichuan Academy of Medical Sciences Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China
Jinyan Yuan
University of Electronic Science and Technology of China
Yi Zhou
Sichuan University
Fei Deng (dengfei_here@163.com)
Chengdu Jinniu District People's Hospital

Research Article

Keywords: antituberculosis drug-associated liver injury, PXR, ALAS1, FOXO1, single nucleotide gene polymorphism, prediction model, column line graph

Posted Date: March 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1450906/v1

Abstract

BACKGROUND: To develop a predictive model for hepatotoxicity due to antituberculosis drugs using a machine learning approach combining general clinical features of the electronic medical record, laboratory indications and genetic features of key genes in the PXR/ALAS1/FOXO1 axis.

METHODS: Using the occurrence of ATDH as the outcome variable, the data were screened for features and model construction based on general clinical features and laboratory test indications, combined with single nucleotide polymorphism characteristics of PXR, FOXO1 and ALAS1 genes, combined with Lasso regression and logistic regression to evaluate the model's goodness of fit, predictive efficacy, discrimination and consistency, and used clinical decision Curve analysis was used to assess the clinical applicability of the models.

RESULTS: The best model had a discriminant efficacy C-index of 0.8164, sensitivity of 34.25%, specificity of 97.99%, positive predictive value of 78.13%, negative predictive value of 87.69%, consistency test Sp=0.896, maximum bias Emax=0.147, and mean bias Eave=0.017. In the validation set performance was close. The clinical decision curve shows the clinical applicability of the prediction model when the prediction risk threshold is between 0.1 and 0.8.

CONCLUSION: The ATDH prediction model was constructed using a machine learning approach, combining general characteristics of the study population, laboratory indications and SNP features of PXR and FOXO1 genes with good fit and some predictive value, and has potential and value for clinical application.

Background

According to the Global Tuberculosis Annual Report 2020, the international and domestic situation of tuberculosis infection remains critical, with China accounting for approximately 8.4% of the global tuberculosis population. One of the side effects of the WHO-recommended first-line anti-tuberculosis regimen is anti-tuberculosis drug induced hepatic injury (ATDH), which occurs at an incidence of 5.0–28.0% and can lead to discontinuation of first-line regimens, treatment failure, and the development and spread of multi-drug resistant nodules [1, 2]. There is a lack of specific clinical symptoms and diagnostic markers for the diagnosis of ATDH. Single nucleotide polymorphisms (SNPs) have been shown to have potential and clinical application as molecular markers of ATDH [3]. However, the results of single nucleotide SNPs do not provide a complete and systematic picture of the relevance of the signaling pathway in which they are located to ATDH [3, 4].

PXR/ALAS axis activation leading to abnormal metabolism of the hepatic heme pathway is one of the possible mechanisms by which combined rifampicin and isoniazid treatment leads to ATDH [5]. PXR is an important factor in activating ALAS1 transcription [6]. The transcriptional balance of the ALAS1 gene is coordinated by a complex combination of signaling pathways. The insulin-sensitive FOXO1 pathway can synergize the transcriptional regulation of ALAS1 by PXR [7, 8]. Previous studies by the group have identified genetic polymorphisms in PXR and FOXO1 that correlate with ATDH susceptibility [9, 10].

The clinical prediction model has the potential to assist in the diagnosis of ATDH by identifying predictors that have a significant impact on outcome through a multifactorial analysis approach, while providing different combinations of candidate predictors to further assess the probability of occurrence of the outcome to assist in clinical practice [11]. The current ATDH prediction models with single sociodemographic factors or partial clinical indications as predictor variables have not been validated for goodness-of-fit, and the robustness and accuracy of their predictive efficacy are yet to be verified. With the development of pharmacogenomics, SNPs as important genetic variable features as predictors are believed to improve model predictive power, and combining multi-gene feature variables can enhance predictive model efficacy [12–14]. Machine learning algorithms can perform big data mining effectively and at high speed to discover potential clinically relevant factors by integrating multidimensional information simultaneously, from which valid information can be extracted for model fitting, thus improving model stability and predictive efficacy in unknown datasets and increasing the applicability of models to assist clinical diagnosis and treatment [15]. Therefore, building predictive models based on general clinical features, laboratory indications and multigene genetic features with machine learning algorithms can lay the foundation for building clinical models that are both stable and individualized. At the same time, the probability of occurrence of disease outcomes predicted by the prediction model can be visualized by column line plots, which is more intuitive and easy to use than conventional prediction formulas[16].

In summary, in this study, for the first time, machine learning algorithms were used simultaneously to establish a visual prediction model of ATDH risk by combining the general clinical characteristics, laboratory indications and genetic characteristics of multiple genes in the PXR/ALAS/FOXO1 axis in 746 patients with confirmed TB, and to evaluate the model with respect to its predictive performance and clinical applicability.

Study Population And Methods

1.1 Study population.

A total of 1060 patients with suspected tuberculosis who attended West China Hospital of Sichuan University from December 2016 to April 2018 were retrospectively included [17]. Inclusion criteria: clear tuberculosis diagnosis according to our tuberculosis diagnostic criteria; use of anti-tuberculosis first-line treatment regimen. Exclusion criteria: unclear TB diagnosis; use of other hepatotoxic drugs; concomitant HIV, HBV, HCV or other immune diseases; failure of follow-up; discontinuation of first-line treatment [18]. This study was authorized by the Ethics Committee of West China Hospital, Sichuan University. All experimental subjects were unrelated Han Chinese population in western China who voluntarily participated in this study and signed the informed consent form. Inclusion criteria for the ATDH group: 1) receiving first-line anti-tuberculosis drug regimen; 2) diagnosis of liver injury in accordance with CTCAEv.5.0 criteria; 3) no other hepatotoxic drugs used 14 days before the diagnosis of ATDH [9, 17]. Inclusion criteria for the non-ATDH group: no liver injury occurred during anti-tuberculosis treatment.

1.2 Selection of target SNPs loci and typing assay

The dbSNP database and the 1000 Genomes database were used, and the haploView software was used to screen the candidate genes PXR, FOXO1 and ALAS1 on the target SNPs, the screening principles were as follows [9, 18]: minimum allele frequency ≥ 20% in the southern Chinese Han population; linkage disequilibrium value of tag SNPs r2 ≥ 0.8; located in the region where the candidate genes are located 200 bp upstream and 300 bp downstream; combined with domestic and international literature, SNPs loci that may be associated with the risk of ATDH development or have potential functional significance were included. Based on the above principles and pre-experimental results, SNPs loci that could be successfully designed with PCR primers and single-base amplification primers and successfully typed were screened and typed using 48-Plex SNPscan® high-throughput SNP typing technology [17, 18].

1.3 Data collection, preprocessing and feature variable screening

The basic medical history of the subjects' clinic was collected by exporting from the HIS system; the relevant laboratory indications were exported in the LIS system. Missing data of < 10% were filled with median for continuous variables and plural for categorical variables; missing data of > 10% were excluded. All study subjects were randomly divided into test and validation datasets according to a 3:2 ratio. Lasso regression was used to initially screen candidate variables, and the smallest penalty coefficient lambada (λ) was selected to construct a subset of candidate variables at $p < 0.2$ [19, 20, 21].

1.4 Construction of prediction models and evaluation of predictive efficacy using test set data

The modeling of candidate variables using the test set was performed using STATA software v15.0, and the goodness of fit of the model was evaluated using Akaike's information Criterion (AIC) [22] [23]. The selection criteria were 1) AIC minimization and 2) candidate variables minimization without affecting predictive efficacy [20, 24]. Covariance and interaction analyses were performed on the included predictors [24] [25]. The performance parameters such as sensitivity, specificity, positive predictive value and negative predictive value were evaluated with discrimination test [19]. We use ROC curves and C-index for model differentiation assessment [23], and use calibration curve plots for consistency assessment [26].

1.5 Validation of prediction model efficacy using validation set data

Using the predictors obtained from the test set model and the corresponding coefficients, model reconstruction and validation of model fit was performed using the validation set data [27].

1.6 Model visualization and clinical application value assessment

A column line graph model was developed to visualize the model [28]. The clinical application value of the model was demonstrated using decision curve analysis, and the strategy with the highest net benefit for a specific threshold probability was considered the best strategy [29] [24].

2.Statistical analysis

SPSS software (version 23.0) was used for data on clinical data and laboratory indications: t-test or ANOVA for quantitative data obeying normal distribution, expressed as mean ± standard deviation (± SD); Mann-Whitney or Kruskal-Wallis nonparametric tests for quantitative data with non-normal distribution, expressed as median and interquartile spacing (M50 [M25-M75]). The chi-square test or logistic regression was used to count the count data [18]. This study jointly used R version 3.6.1 software to screen potential predictors through Lasso regression as well as SPSS version 23.0 software using one-way logistic regression analysis with $p < 0.20$ as the judgment threshold for inclusion of predictors. Multi-factor analysis was performed by STATA.vesion 14 software using generalized linear model logistic regression stepwise selection method, and the model was constructed using the minimum value of AIC and the minimum number of predictors as criteria. ROC curve analysis was used to evaluate the predictive model discrimination using C-index as a criterion. The Hosmer-lemeshow test was used to evaluate the consistency of the prediction model in the validation set data, with $p > 0.05$ as the reference criterion, and the Unreliability test (U test) corrected curve analysis was used to evaluate the consistency of the model fit, with $p > 0.05$ as the reference criterion [30]. The prediction model was visualized using column line graphs. The clinical application value of the prediction model was analyzed using decision curves.

Results

3.1 Basic information of the study population

A total of 746 study subjects (118 in the ATDH group and 628 in the non-ATDH group) were included in this study. There was no statistical difference between the two groups in the proportion of gender, age, and living habits. However, the proportion of patients presenting with febrile symptoms was significantly lower in the ATDH group than in the non-ATDH group, as shown in Table 1.

3.2 Indications of clinical laboratory tests in the study population

Patients in the ATDH group had increased TBIL, indirect bilirubin, AST, ALT, alkaline phosphatase and glutamyl transferase levels and lower uric acid levels relative to those in the non-ATDH group (all $p < 0.05$), as shown in Table 2.

3.3 Loci typing results for the target SNPs in the study population

T allele carriers at rs3814055 of the PXR gene had a reduced relative risk of ATDH compared to C allele carriers [9]. Carriers of the rs2755237 locus C allele of the FOXO1 gene had a reduced relative risk of ATDH relative to carriers of the A allele and carriers of the T allele of the rs4435111 locus relative to carriers of the C allele. The gene frequencies of candidate SNPs for the ALAS1 gene did not differ between the two groups.

3.4. modelling of ATDH risk prediction

3.4.1 Model predictor screening

Lasso regression in the machine learning algorithm was used to screen the pre-treated 98 characteristic variables, showing that the optimal subset of non-zero coefficient variables for inclusion in the model was 36 at the minimum value of 10-fold cross-validation error $\lambda = 0.0074528$, and the coefficients of the remaining variables were reduced to zero, as shown in S 1 and 2.

3.4.2 Identification of candidate predictors using one-way logistic regression

As shown in Table 3, there are 12 candidate feature variables that are statistically different in the test set, respectively. There are 9 corresponding candidate feature variables in the validation set. The characteristic variables that were statistically different in both groups were total bile acid, glutamic aminotransferase, glutamic oxalacetic aminotransferase, and uric acid.

3.4.3 Adjustment for model confounders

There was moderate strength covariance p = 0.616 for ALT and AST and no multicollinearity between the remaining 15 candidate variables two by two with a maximum p-value of 0.26. Rs3814055 and rs4435111 had an interaction effect on the outcome variable ATDH occurrence (p = 0.001). No interactions were detected between the other 15 variables, all p > 0.05.

3.4.4 Test set model building and optimization

The 17 candidate predictors were modeled in different ways, and the screening p-values and AIC and BIC are shown in Table 4. Model 6 incorporated 5 variables with an AIC of 320.50, model 8 incorporated 9 variables with an AIC of 312.68, and model 9 incorporated 8 variables with an AIC of 312.44. Comparison of model 6, model 8 and model 9 revealed that model 6 and model 8 were different, and although model 6 incorporated fewer variables, its predictive efficacy was reduced (using STATA software's lrtest test command, p < 0.05). In contrast, there was no difference in predictive efficacy between model 8 and model 9, but model 9 incorporated fewer variables (using the lrtest test command of STATA software, p > 0.05), thus model 9 was considered the best model with the characteristics of incorporated variables as shown in Table 5.

3.4.5 Test set model predictive efficacy analysis

The model had a discriminant C-index of 0.816, a sensitivity of 34.25%, a specificity of 97.99%, a positive predictive value of 78.13%, and a negative predictive value of 87.69%, as shown in S 3. The model consistency test had Sp = 0.896, maximum offset Emax = 0.147 and mean offset Eave = 0.017, as shown in S 4.

3.4.6 Validation set model building and effectiveness analysis

Logistic regression models were recreated in the validation set data data summary using the regression coefficients from the test set model.

Odds (ATDH) = 1/(1 + exp(-(-3.661122 + 0.7491207 × Fever + 0.0676586 × Alb-0.0023242 × Uric + 0.1458457 × Monocyte % + 0.050343 × AST + 0.0662291 × ALT − 1.373078 × rs4435111−0.5698482 × rs3814055))).

The fit of this model was consistent with the model constructed from the test set data (Hosmer-Lemeshow test p = 0.4636). Validation set the model ROC curve analysis discrimination C-index 0.7196, specificity 97.77%, negative predictive value 86.21%, sensitivity 15.15%, positive predictive value 55.56%, as shown in S 5; calibration curve validation maximum offset Emax = 0.101, mean offset Eave = 0.009, p = 0.929, as shown in S 6.

3.4.7 Building the column line graph model

The column line graph was established according to this prediction model, and the genotypes of rs3814055 were stratified because of the interaction between rs3814055 and rs4435111. Because the different genotypes of rs3814055 and rs4435111 had non-equal predicted risk of ATDH, the different genotypes were set according to dummy variables. The column line graph model is shown in S 7, with predicted probabilities between approximately in the range of in the range of 0.1–0.7 for total integrals between 170–210.

3.4.8 DCA effects analysis of the prediction model

The clinical decision curve for the ATDH prediction model is shown in S 8. The model has value for clinical use when the risk threshold ranges between 0.1 and 0.8.

Discussion

In this study, the ATDH prediction model was constructed by using machine learning algorithms to screen eight predictors in terms of general clinical characteristics, laboratory indications and genetic characteristics variables. The construction process was carried out strictly with reference to the statement of clinical prediction models using three steps: developing the prediction model, validating the prediction model, and studying the clinical significance of the model [31]. The model had moderate specificity, discrimination, consistency and clinical application, however, the sensitivity needs to be improved.

In this study, Lasso regression, a machine learning algorithm, was used for pre-screening of model feature variables. Lasso regression is beneficial in constructing prediction models to satisfy variance trade-offs while integrating a large amount of data in

different dimensions. It has the characteristics of fast analysis, stability and easy interpretation of results compared to the conventional logistic regression step-by-step processing [32]. Therefore, this study first used Lasso regression for data pre-screening and further used one-way logistic regression to filter out 17 candidate variables [21].

The group used the principle of multivariate logistic regression in the test set data, with the lowest AIC value and the least number of predictors as the selection criteria for the optimal model [33]. The included predictors were fever, rs3814055, rs4435111, albumin, ghrelin, glutamic aminotransferase, uric acid, and monocyte percentage.

Visualization of the optimized model revealed that alanine transaminase (ALT), aspartate transaminse (AST), albumin, monocyte percentage, and fever were all independent predictors of ATDH, suggesting that the basal liver function status, immune status and ATDH susceptibility in TB patients were associated. Meanwhile, the column line graph visualized the interaction between rs3814055 and rs4435111: when rs3814055 was CC genotype, the risk of ATDH were significantly increased, while TT genotype of rs4435111 was able to relatively reduce the risk of ATDH. When rs3814055 was TT genotype, the overall risk of ATDH occurred decreased, while TT genotype of rs4435111 relatively increased the risk of ATDH. Rs4435111's TT genotype was in the genetics section a factor in the reduced risk of ATDH occurrence, however the column line graph model showed that its score value for predicting the probability of ATDH risk was influenced by the genotype at the rs3814055 locus. Analysis of possible reasons for this is as follows: 1) both rs3814055 and rs4435111 have relatively few TT genotypes, leading to bias in data from small samples; and 2) there is a complex higher-order and second-order multiplicative interaction of the PXR gene with the FOXO1 gene. The rs3814055 genotype interfered with the efficacy of the latter assessment due to the greater weighting of the rs3814055 genotype on the effect of ATDH susceptibility.

The model had a C-index = 0.8164 for the test set's discriminant test, The consistency test p = 0.896, Emax = 0.147, and Eave = 0.017. This suggest that both the model's discriminant and consistency were good. In order to avoid overfitting of the model due to random and systematic errors in the cross-validation data in different training data sets, the model fit needs to be validated in the validation set data to prevent the increase in variance caused by overfitting. It was shown that the fit of the model constructed from the validation set data was consistent with that of the model constructed from the test set data, and the discrimination had a moderate strength of discrimination, indicating that the use of Lasso regression was effective in preventing model overfitting from causing fit contraction in the new sample set. Further clinical decision curve analysis of the model revealed that when the high risk threshold was between 0.1 and 0.8, it suggested that the model was of good value for clinical use.

However, the model's sensitivity was 34.25% and 15.15% in the test and validation sets, respectively, and its specificity was 97.99% and 97.77%, respectively, with positive predictive values of 78.13% and 55.56% and negative predictive values of 87.69% and 86.21%, also suggesting that its predictive sensitivity needs to be improved.

The possible reasons for the good predictive specificity and poor sensitivity of the model are as follows: 1) Low incidence of ATDH. In this study, there were 118 cases in the ATDH group and 628 cases in the non-ATDH group, and the incidence of ATDH was 15.81%. The group randomly divided all TB patients into the test set (91 cases in the ATDH group) and the validation set (37 cases in the ATDH group) in a 3:2 ratio, and their ATDH incidence was 18.57% and 14.45%, corresponding to a sensitivity of 34.25% and 15.15%, respectively. The low incidence of ATDH in the constructed model data may be one of the important reasons for the poor sensitivity of the model.2) Lack of strong predictors. The predictors selected by Lasso regression and one-way logistic regression for model inclusion factors were general clinical features (fever), routine laboratory indications (ALT, AST, Alb, monocyte percentage) and genetic indications (genotype of rs3814055 and rs4435111), respectively. Although these predictors are objective tests, they are all relevant markers derived from the mechanisms of ATDH occurrence and not specific markers.

Therefore, based on our established prediction model for ATDH, it can be concluded that 1) the machine learning algorithm Lasso regression helps to simultaneously perform a large number of candidate variables screening, meet the requirements of variance trade-off by bootstrap self-sampling and cross-validation, and avoid overfitting; 2) SNPs are promising predictors, and combining multi-gene SNPs features over single-gene SNPs to build prediction models can improve predictive efficacy and clinical applicability; 3) Simultaneous modeling of multi-gene SNPs requires consideration of the impact of interactions on model predictive efficacy. Further research directions should also be validated in a larger and different population, while adding as many key genes or clinical data as possible to increase the sensitivity of the model.

Abbreviations

ATDH: anti-tuberculosis drug induced hepatic injury

SNPs: Single nucleotide polymorphisms

AIC: Akaike's information Criterion

M50 [M25-M75]: median and interquartile spacing

U test: Unreliability test

ALT: alanine transaminase

AST: aspartate transaminse

Declarations

- **Ethical Approval and Consent to participate**

Not applicable.

- Consent for publication

Not applicable.

- Availability of supporting data

Not applicable.

- Competing interests

The authors declare that they have no competing interests.

- Funding

Not applicable.

- Authors' contributions

JZ and WZ conceived the study and designed experimental procedures. JZ, ZW, PL and WX performed the experiments. WZ, XL, and JY analyzed the data. JZ contributed to reagents and materials. JZ and WZ wrote the original draft. YZ and FD reviewed the manuscript. All authors read and approved the final manuscript.

- Acknowledgements

Not applicable.

Conflict of Interest Statement: The authors have declared that no competing interest exists.

References

1. Chakaya J, Khan M, Ntoumi F, Aklillu E, Fatima R, Mwaba P, *et al.* Global Tuberculosis Report 2020 - Reflections on the Global TB burden, treatment and prevention efforts. International journal of infectious diseases: IJID : official publication of the International Society for Infectious Diseases 2021.
2. Bao Y, Ma X, Rasmussen T, Zhong X. Genetic Variations Associated with Anti-Tuberculosis Drug-Induced Liver Injury. Current pharmacology reports 2018, 4(3): 171–181.

3. Bao Y, Ma X, Rasmussen TP, Zhong XB. Genetic Variations Associated with Anti-Tuberculosis Drug-Induced Liver Injury. Curr Pharmacol Rep 2018, 4(3): 171–181.
4. Huang YS. Recent progress in genetic variation and risk of antituberculosis drug-induced liver injury. J Chin Med Assoc 2014, 77(4): 169–173.
5. Lyoumi S, Lefebvre T, Karim Z, Gouya L, Puy H. PXR-ALAS1: a key regulatory pathway in liver toxicity induced by isoniazid-rifampicin antituberculosis treatment. Clin Res Hepatol Gastroenterol 2013, 37(5): 439–441.
6. Podvinec M, Handschin C, Looser R, Meyer UA. Identification of the xenosensors regulating human 5-aminolevulinate synthase. Proceedings of the National Academy of Sciences 2004, 101(24): 9127–9132.
7. Fraser DJ, Zumsteg A, Meyer UA. Nuclear receptors constitutive androstane receptor and pregnane X receptor activate a drug-responsive enhancer of the murine 5-aminolevulinic acid synthase gene. J Biol Chem 2003, 278(41): 39392–39401.
8. THUNELL S. Genomic Approach to Acute Porphyria. Physiol Res 2006, 55(2).
9. Zhang J, Zhao Z, Bai H, Wang M, Jiao L, Peng W, *et al.* Genetic polymorphisms in PXR and NF-kappaB1 influence susceptibility to anti-tuberculosis drug-induced liver injury. PLoS One 2019, 14(9): e0222033.
10. Zhang J, Jiao L, Song J, Wu T, Bai H, Liu T, *et al.* Genetic and Functional Evaluation of the Role of FOXO1 in Antituberculosis Drug-Induced Hepatotoxicity. Evidence-Based Complementary and Alternative Medicine 2021, 2021: 1–13.
11. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014, 35(29): 1925–1931.
12. Pontual Y, Pacheco V, Monteiro S, Quintana M, Costa M, Rolla V, *et al.* gene polymorphism associated with clinical factors can predict drug-resistant tuberculosis. Clinical science (London, England: 1979) 2017, 131(15): 1831–1840.
13. Mushiroda T, Yanai H, Yoshiyama T, Sasaki Y, Okumura M, Ogata H, *et al.* Development of a prediction system for anti-tuberculosis drug-induced liver injury in Japanese patients. Hum Genome Var 2016, 3(16): 14–18.
14. Chamorro JG, Castagnino JP, Aidar O, Musella RM, Frias A, Visca M, *et al.* Effect of gene-gene and gene-environment interactions associated with antituberculosis drug-induced hepatotoxicity. Pharmacogenet Genomics 2017, 27(10): 363–371.
15. Mahomed S, Padayatchi N, Singh J, Naidoo K. Precision medicine in resistant Tuberculosis: Treat the correct patient, at the correct time, with the correct drug. Journal of Infection 2019, 4(54): 1–8.
16. Guo1 B-L, Shao-Jia Lin WM, Ouyang F-S, Huang X-Y, Yang S-M, Ouyang L-Z, *et al.* Development of a preprocedure nomogram for predicting contrast-induced acute kidney injury after coronary angiography or percutaneous coronary intervention. Oncotarget 2017, 8(44): 75087–75093.
17. Zhang J, Zhao Z, Bai H, Jiao L, Wu Q, Wu T, *et al.* The Variant at TGFBRAP1 but Not TGFBR2 Is Associated with Antituberculosis Drug-Induced Liver Injury. Evid Based Complement Alternat Med 2019, 23(1): 1–10.
18. Li Yingyu. Clinical characteristics of adverse reactions due to anti-tuberculosis drugs and their association with genetic polymorphisms. 2018.
19. Huang Y, Liang C, He L, Tian J, Lian C, Chen X, *et al.* Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. JOURNAL OF CLINICAL ONCOLOGY 2016, 34(10): 109.
20. Stone GW, Maehara A, Lansky AJ, de Bruyne B, Cristea E, Mintz GS,*et al.* A prospective natural-history study of coronary atherosclerosis. N Engl J Med 2011, 364(3): 226–235.
21. Kang SJ, Cho YR, Park GM, Ahn JM, Han SB, Lee JY, *et al.* Predictors for functionally significant in-stent restenosis: an integrated analysis using coronary angiography, IVUS, and myocardial perfusion imaging. JACC Cardiovasc Imaging 2013, 6(11): 1183–1190.
22. Akaike H. Data analysis by statistical models. No To Hattatsu 1992, 24(2): 127–133.
23. Yan Ruohua. Establishment and validation study of a lung function assessment model for initial screening of chronic obstructive pulmonary disease in China. Doctoral dissertation, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, 2019.
24. Jaddoe VW, de Jonge LL, Hofman A, Franco OH, Steegers EA, Gaillard R. First trimester fetal growth restriction and cardiovascular risk factors in school age children: population based cohort study. Bmj 2014, 34(8): 14–25.

25. Qiao F, Fu K, Zhang Q, Liu L, Meng G, Wu H, *et al.* The association between missing teeth and non-alcoholic fatty liver disease in adults. J Clin Periodontol 2018, 45(8): 941–951.
26. Coutant C, Olivier C, Lambaudie E, Fondrinier E, Marchal F, Guillemin F, *et al.* Comparison of models to predict nonsentinel lymph node status in breast cancer patients with metastatic sentinel lymph nodes: a prospective multicenter study. J Clin Oncol 2009, 27(17): 2800–2808.
27. Gao Jiali. Value of chest CT combined with serum tumor markers in the differential diagnosis of benign and malignant pulmonary nodules. Professional master's thesis, Zhengzhou University 2019.
28. Chen Juan. Application of column line graph model in slow plus acute liver failure associated with viral hepatitis B. Master's thesis, Fujian Medical University 2016.
29. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak 2008, 8: 53.
30. Yang DX. Construction and validation of an intraoperative risk prediction line chart for laparoscopic left hepatectomy for hepatobiliary stone disease. Dissertation for Professional Degree Master's Degree, Nanchang University, 2019.
31. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015, 162(1): W1-73.
32. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical SocietySeries B (Methodological 1996, 58(1): 267–288.
33. Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychol Methods 2012, 17(2): 228–243.

Tables

Table 1 Basic clinical characteristics of the study subjects

Basic Characteristics	ATDH (n = 118)	Non-ATDH(n = 628)	*p*
Age ±SD (years)	40.92 ± 15.72	42.85 ± 18.44	0.285
Gender Male/Female	69(58.47)/49(41.53)	375(59.71)/253(40.29)	0.801
Smoking Yes/No	35(29.66)/83(70.34)	192(30.57)/436(69.43)	0.843
Drinking alcohol yes/no	32(27.12)/86(72.88)	141(22.45)/487(77.55)	0.270
General symptoms n (%)			
Fever yes/no	63(20.30)	247(79.7)	0.004
Weight loss yes/no	34(13.10)	226(86.9)	0.133
Nocturnal night sweats yes/no	29(14.80)	167(85.20)	0.648
Loss of appetite yes/no	45(17.10)	218(82.90)	0.475
Fatigue yes/no	31(17.90)	142(82.10)	0.387
Local infection symptoms (%)			
Appearances	88(16.90)	432(83.10)	0.209
Disappearances	30(13.30)	196(86.70)	

Table 2 Indications for clinical laboratory tests in the study population

Baseline values for laboratory test results (±SD or M(P25-P75)	ATDH group (n = 118)	Non-ATDH group (n = 628)	*p*
Red blood cell count ($\times 10^{12}$/L)	4.31 ± 0.74	4.28 ± 0.68	0.481
Haemoglobin (g/L)	122.87 ± 22.11	122.06 ± 20.58	0.717
Erythrocyte pressure (%)	0.38 ±0.06	0.36 ±0.06	0.069
Platelet count($\times 10^9$/L)	236.50(184.00-321.75)	232.50(172.75-297.25)	0.134
White blood cell count($\times 10^9$/L)	6.57(4.99-7.96)	6.51(5.17-8.44)	0.761
Absolute value of neutrophils($\times 10^9$/L)	5.23 ± 2.89	5.10 ± 2.73	0.631
Absolute value of lymphocytes($\times 10^9$/L)	1.29 ± 0.79	1.26 ± 0.62	0.625
Absolute value of monocytes($\times 10^9$/L)	0.55 ± 0.29	0.50 ± 0.25	0.099
Neutrophils(%)	70.49 ± 11.50	70.13 ± 11.54	0.760
Lymphocytes(%)	16.25(12.58-25.58)	17.5(12.18-25.68)	0.527
Monocytes(%)	7.74 ± 2.62	7.30 ± 2.37	0.077
Total bilirubin(μmol/L)	10.05(7.50-14.13)	8.70(6.30-12.10)	0.002
Direct bilirubin(μmol/L)	3.55(2.38-5.60)	3.45(2.50-5.40)	0.126
Indirect bilirubin(μmol/L)	5.70(3.98-7.95)	4.80(3.40-7.03)	0.049
ALT(IU/L)	28.00(15.75-38.00)	15.00(10.00-21.00)	<0.001
AST(IU/L)	27.00(20.00-34.00)	19.50(16.00-25.00)	<0.001
Total protein(g/L)	69.42 ± 8.42	68.82 ± 9.15	0.508
Albumin(g/L)	38.64 ± 7.35	37.89 ± 6.90	0.248
Globulin(g/L)	30.78± 6.65	30.93 ± 7.02	0.829
Glucose(mmol/L)	5.15(4.64-5.95)	5.14(4.71-5.89)	0.410
Urea(mmol/L)	3.92(2.90-5.24)	4.05(3.15-5.30)	0.299
Creatinine(umol/L)	57.50(47.78-67.00)	60.45(49.00-73.20)	0.601
Serum cystatin C(mg/L)	0.91(0.81-1.04)	0.92(0.79-1.06)	0.975
Uric acid(umol/L)	291.29± 125.98	331.51± 155.30	0.008
Triglycerides(mmol/L)	0.99(0.81-1.31)	1.06(0.80-1.43)	0.469
Cholesterol(mmol/L)	3.96±1.206	3.96±1.058	0.966
High-density lipoprotein(mmol/L)	1.12(0.85-1.48)	1.08(0.82-1.41)	0.811
Low-density lipoprotein(mmol/L)	2.20(1.79-2.72)	2.21(1.69-2.77)	0.575
Alkaline phosphatase(IU/L)	85.50(68.50-106.00)	79.00(64.00-98.00)	0.021
Glutamyl transferase(IU/L)	42.50(26.00-78.00)	29.00(19.00-48.00)	<0.001
C-reactive protein(mg/L)	9.74(2.30-39.23)	12.25(2.67-37.43)	0.961
Blood sedimentation(mm/h)	38.50(20.50-63.00)	33.50(14.75-64.00)	0.173

Table 3 Baseline levels of the 36 characteristic variables in the test and validation sets after Lasso screening

Candidate feature variables		Test set（n=490）					Validation set（n=256）				
		Non-ATDH group（n=409）		ATDH group（n=91）		*p*	Non-ATDH group（n=219）		ATDH group（n=37）		*p*
Gender (m/f; n, %)		246（60.1）	163（39.9）	50（54.9）	41（45.1）	0.409	129(58.9)	90(41.1)	24(64.9)	13(35.1)	0.588
Alcohol consumption (yes/no; n, %)		282（68.9）	127（31.1）	61（67.0）	30（33.0）	0.710	162(74.0)	57(26.0)	25(67.6)	12(32.4)	0.427
Fever (yes/no; n, %)		232（56.7）	177（43.3）	37（40.7）	54（59.3）	0.007	108(49.3)	111(50.7)	19(51.4)	18(48.6)	0.860
Weight loss (yes/no; n, %)		233（57.0）	176（43.0）	57（62.6）	34（37.4）	0.349	126(57.5)	93(42.5)	28(75.7)	9(24.3)	0.045
Decreased appetite (yes/no; n, %)		243（59.4）	166（40.6）	51（58.8）	40（41.2）	0.558	122(55.7)	97(44.3)	23(62.2)	14(37.8)	0.480
Fatigue (yes/no; n, %)		287（70.2）	122（29.8）	63（69.2）	28（30.8）	0.900	153(69.9)	66(30.1)	24(66.7)	12(33.3)	0.700
rs353556	11	116	28.4	22	27.2	0.886	57	26	9	24.3	0.544
（n（%）	22	207	50.6	40	49.4		120	54.8	18	48.6	
	33	86	21	19	23.5		42	19.2	10	27.0	
rs3852071（n（%）	11	13	3.2	0	0.0	0.267	3	1.4	1	2.7	0.180
	22	111	27.3	22	27.5		67	30.6	6	16.2	
	33	283	69.5	58	72.5		149	68	30	81.1	
rs352169（n（%）	11	174	42.6	27	33.8	0.075	80	36.5	17	45.9	0.334
	22	195	47.8	39	48.8		103	47	17	45.9	
	33	39	9.6	14	17.5		36	16.4	3	8.1	
rs2755237（n（%）	11	191	46.9	53	66.3	0.007	101	46.3	15	40.5	0.373
	22	192	47.2	24	30.0		101	46.3	21	58.6	
	33	24	5.9	3	3.8		16	7.3	1	2.7	
rs2701891（n（%）	11	224	54.8	39	48.1	0.009	126	57.5	22	59.5	0.202
	22	160	39.1	29	35.8		76	34.7	15	40.5	
	33	25	6.1	13	16.0		17	7.8	0	0.0	
rs3751436（n（%）	11	149	36.5	28	35.4	0.711	80	36.5	10	27	0.481
	22	206	50.5	38	48.1		108	49.3	22	59.5	
	33	53	13	13	16.5		31	14.2	5	13.5	
rs4435111（n（%）	11	249	60.9	63	78.8	0.005	117	53.4	20	4.1	0.991
	22	144	35.2	17	21.3		89	40.6	15	40.5	
	33	16	3.9	0	0.0		13	5.9	2	5.4	
rs7325594（n（%）	11	51	12.5	13	16.0	0.506	31	14.2	5	13.5	0.329
	22	218	53.4	45	55.6		109	49.8	23	62.2	
	33	139	34.1	23	28.4		79	36.1	9	24.3	

rs3814055 （n（%）	11	242	59.6	56	61.2	0.240	131	59.8	28	75.7	0.114
	22	137	33.7	22	27.2		76	34.7	9	24.3	
	33	27	6.7	3	3.7		12	5.5	0	0.0	
rs56967099 （n（%）	11	111	27.3	24	30.0	0.841	60	27.5	9	24.3	0.901
	22	189	46.4	37	46.3		104	47.7	19	51.4	
	33	107	26.3	19	23.8		54	24.8	9	24.3	
rs13059232 （n（%）	11	157	38.4	34	42.5	0.674	80	36.5	14	37.8	0.939
	22	196	47.9	34	42.5		107	48.9	17	45.9	
	33	56	13.7	12	15.0		32	14.6	6	16.2	
rs4688040 （n（%）	11	149	37.3	34	42.5	0.532	82	38.3	13	36.1	0.933
	22	197	49.4	34	42.5		101	47.2	17	47.2	
	33	53	13.3	12	15.0		31	14.5	6	16.7	
rs6785049 （n（%）	11	163	40.1	39	48.1	0.401	94	42.9	15	40.5	0.802
	22	187	46.1	33	40.7		95	43.4	18	48.6	
	33	56	13.8	9	11.1		30	13.7	4	10.8	
rs3732360 （n（%）	11	129	31.5	25	30.9	0.675	76	34.7	18	48.6	0.220
	22	206	50.4	38	46.9		108	49.3	13	35.1	
	33	74	18.1	18	22.2		35	16	6	16.2	

Platelets（10^9/L）	231（171-296.5）	244（185.5-322.5）	0.0549	235（173-293）	221（181.5-276）	0.9589
Percentage of neutrophils（%）	71.7（62-78.3）	70.6（61.28-76.53）	0.6908	70（64-79）	74.5（65.35-82.65）	0.2106
Percentage of monocytes (%)	7.1（5.75-8.9）	8.3（5.68-9.45）	0.1136	6.9（5.8-8.8）	7.4（5.05-8.84）	0.5629
Absolute monocyte values （10^9/L）	0.47（0.35-0.64）	0.49（0.36-0.67）	0.341	0.45（0.32-0.6）	0.46（0.36-0.76）	0.2338
Total bile acids （μmol/L）	8.7（6.4-12.225）	9.8（7.6-14.55）	0.0087	8.8（6.4-12.2）	10.4（6.6-14.7）	0.1414
Direct bilirubin (μmol/L)	3.5（2.5-5.4）	3.6（2.3-6.6）	0.3726	3.5（2.5-5.45）	3.9（2.9-7.55）	0.1374
Glutamic-pyruvic transaminase（IU/L）	14（10-20）	27（13.5-38）	<0.0001	16（10-23）	27（17-38）	<0.0001
Glutathane transaminase（IU/L）	19（15-25）	27（19-34）	<0.0001	21（17-26.25）	25（21-33）	0.0012
Albumin（g/L）	38.5（33.2-43.1）	38.7（34.4-43.6）	0.2744	38.95（33.1-43.325）	37.8（30.4-46.85）	0.866
Glucose（mmol/L）	4.72（5.16-5.97）	4.61（4.93-5.64）	0.0735	5.05（4.65-5.67）	5.37（4.57-6.24）	0.2974
Creatinine（umol/L）	61.1（49-74）	57.4（47.5-69）	0.2144	59（50-71）	61（53-73）	0.3446
Uric acid (umol/L)	306.7（225.5-403）	273（203-393）	0.1429	292（224-417）	267.35（185.25-350.25）	0.0392

Triglycerides (mmol/L)	1.04（0.78-1.41)	1.02（0.81-1.47)	0.9415	1.01（0.78-1.39)	0.94（0.83-1.13)	0.1614
Alkaline phosphatase (IU/L)	78（64-96.25)	84（72.5-104)	0.0254	80（63.75-99)	80（65.5-108.5)	0.3916

Table 4 Multiple models using multivariate logistic regression for comparison

models	Construction method	Inclusion of variables	Screening p-value	number of variables	AIC	BIC
Model1	entry into law	All variables	/	17	325.71	406.32
Model2	entry into law (dummy variable)	All variables	/	17	327.26	422.75
Model 3	entry into law (dummy variable)	rs4435111、rs3814055、Monocyte%、PLT、ALT、AST、Uric、ALP、TBIL、DBIL、Alb、Glu、TG	/	13	350.91	408.60
Model 4	entry into law(dummy variable)	rs4435111、rs3814055、Monocyte%、ALT、AST、Uric、Tbil	/	7	352.18	389.40
Model 5	Stepwise method	All variables	0.2	9	312.44	352.74
Model 6	Stepwise method	All variables	0.05	5	320.50	344.68
Model 7	Stepwise method	All variables	0.3	11	313.42	361.78
Model 8	Stepwise method	All variables	0.2	9	312.44	352.74
Model 9	Stepwise method	All variables	0.05	8	312.68	348.95
Model 10	Stepwise method	All variables	0.3	13	315.68	372.11
Model 11	entry into law (dummy variable/interaction)	Fever、rs4435111、rs3814055、Monocyte%、ALT、AST、Uric、Tbil	/	10	315.01	359.16

AIC：Akaike's information Criterion；BIC：Baysian information Criterion

Table 5 Variables and characteristics that were eventually included in the model

characteristic variable	ß	OR	95% CI		*p*
fever	0.7491207	2.115	0.148	1.349	0.015
rs4435111*	-1.373078	0.253	-2.090	-0.65	<0.001
rs3814055*	-0.5692482	0.565	-1.09	-0.04	0.033
albumin	0.0676586	1.070	0.017	0.117	0.008
Glutamic-pyruvic transaminase	0.0662291	1.068	0.036	0.096	<0.001
Glutathane transaminase	0.0503438	1.051	0.004	0.096	0.032
uric acid	-0.0023242	0.997	-0.005	0.00015	0.036
Percentage of monocytes	0.1458457	1.157	0.028	0.262	0.014

*：Both are set up according to dummy variables and modeled in layers according to interactions

Supplementary Files

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

- S1.png
- S2.png
- S3.png
- S4.png
- S5.png
- S6.png
- S7.png
- S81.png