84842_Auto_Edited.docx

WORD COUNT	5537	TIME SUBMITTED	08-MAY-2023 02:09PM
		PAPER ID	99470787

Name of Journal: *World Journal of Gastroenterology*

Manuscript NO: 84842

Manuscript Type: ORIGINAL ARTICLE

Retrospective Study

Comparison and development of machine learning for thalidomide-induced peripheral neuropathy prediction of refractory Crohn's disease in Chinese population

Predicting neurotoxicity using machine learning

Jing Mao, Kang Chao, Fulin Jiang, Xiaoping Ye, Ting Yang, Pan Li, Xia Zhu, Pin-Jin Hu, Baijun Zhou, Min Huang, Xiang Gao, Xue-Ding Wang

Abstract

BACKGROUND

Thalidomide is an effective treatment for refractory Crohn's disease (CD). However, thalidomide-induced peripheral neuropathy (TiPN), which has a large individual variation, is a major cause of treatment failure. TiPN is rarely predictable and recognized, especially in CD. It is necessary to develop a risk model to predict TiPN occurrence.

AIM

To develop and compare a predictive model of TiPN using machine learning based on comprehensive clinical and genetic variables.

METHODS

A retrospective cohort of 164 CD patients from January 2016 to June 2022 was used to establish the model. The National Cancer Institute Common Toxicity Criteria Sensory Scale (version 4.0) was used to assess TiPN. With 18 clinical features and 150 genetic variables, five predictive models were established and evaluated by the confusion matrix receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), specificity, sensitivity (recall rate), precision, accuracy, and F1 score.

RESULTS

The top-ranking five risk variables associated with TiPN were interleukin-12 rs1353248 [P = 0.0004, odds ratio (OR) 8.983, 95% confidence interval (CI) 2.497–30.90), dose (mg/d, *P* = 0.002), brain-derived neurotrophic factor (BDNF) rs2030324 (*P* = 0.001, OR 3.164, 95%CI 1.561–6.434), BDNF rs6265 (*P* = 0.001, OR 3.150, 95%CI 1.546–6.073) and BDNF rs11030104 (*P* = 0.001, OR 3.091, 95%CI 1.525–5.960). In the training set, gradient boosting decision tree (GBDT), extremely random trees (ET), random forest, logistic regression and extreme gradient boosting (XGBoost) obtained AUROC values > 0.90 and AUPRC > 0.87. Among these models, XGboost and GBDT obtained the first two highest AUROC (0.90 and 1), AUPRC (0.98 and 1), accuracy (0.96 and 0.98), precision (0.90 and 0.95), F1 score (0.95 and 0.98), specificity (0.94 and 0.97), and sensitivity (1). In the validation set, XGBoost algorithm exhibited the best predictive performance with the highest specificity (0.857), accuracy (0.818), AUPRC (0.86) and AUROC (0.89). ET and GBDT obtained the highest sensitivity (1) and F1 score (0.8). Overall, compared with other state-of-the-art classifiers such as ET, GBDT and RF, XGBoost algorithm not only showed a more stable performance, but also yielded higher ROC-AUC and PRC-AUC scores, demonstrating its high accuracy in prediction of TiPN occurrence.

CONCLUSION

The powerful XGBoost algorithm accurately predicts TiPN using 18 clinical features and 14 genetic variables. With the ability to identify high-risk patients using single nucleotide polymorphisms, it offers a feasible option for improving thalidomide efficacy in CD patients.

Key Words: Thalidomide-induced peripheral neuropathy; Refractory Crohn’s disease; Neurotoxicity prediction models; Machine learning; Gene polymorphisms.

Mao J, Chao K, Jiang F, Ye X, Yang T, Li P, Zhu X, Hu PJ, Zhou B, Huang M, Gao X, Wang XD. Comparison and development of machine learning for thalidomide-induced peripheral neuropathy prediction of refractory Crohn’s disease in Chinese population. *World J Gastroenterol* 2023; In press

Core Tip: Thalidomide-induced peripheral neuropathy (TiPN) is a life-threatening condition in Crohn's disease and has a high incidence in Asia. However, there are no effective medical interventions for TiPN. Here, we established a predictive model using machine learning and identified genes closely related to TiPN occurrence. We have found that XGboost algorithm can sensitively identify patients who are prone to TiPN, which is useful for doctors to adjust the thalidomide therapy.

INTRODUCTION

Thalidomide is widely used in refractory Crohn's disease (CD) patients, with a remission rate of 40%–70% [1, 2]. However, the clinical application of thalidomide is limited by its side effects, especially peripheral neuropathy. A large individual difference (20%–75%) was found in the incidence of peripheral nerve lesions [3]. If neurotoxicity occurs during the treatment, reduction or cessation is needed to avoid further neurotoxicity. This may lead to treatment failure [4]. The nerve lesions after withdrawal may worsen for several months, recovery can be slow and incomplete, and the resulting neurotoxicity profoundly affects quality of life [5]. Hence, developing a predictive model for thalidomide-induced peripheral neuropathy (TiPN) and identifying related factors that can accurately predict peripheral neuropathy are important.

Thalidomide is a small molecule with immunomodulatory activity [4]. Numerous clinical studies have confirmed that thalidomide is effective for treatment of CD, especially in patients with hormone intolerance, lack of efficacy of azathioprine/6-mercaptopurine, and biological treatment failure [6, 7], It brings hope to patients with refractory CD, which can reduce economic and healthcare costs because of its low price from the perspective of pharmacoeconomics. A retrospective multicenter observational study showed that the proportion of adults with refractory CD from whom thalidomide was withdrawn because of toxicity alone was up to 46% at 2 years [7], which affected maintenance of remission in nearly half of patients treated with thalidomide. TiPN may appear primarily as sensory peripheral nerve lesions and may differ in symptoms, including paralysis, sensory disturbance, sensory abnormalities, hyperalgesia and severe pain [8]. The empiric identification of agents and interventions to mitigate TiPN has been disappointing, and presently there is no intervention available for prevention except dose management. Although electrophysiological monitoring has been used in patients prescribed thalidomide, it provides no clear benefit for the occurrence of neuropathy compared to clinical assessment [3, 9].

Few researchers have studied risk factors related to TiPN [10, 11], and most studies have been conducted in chemotherapy patients. Szczyrek *et al* used serum brain- derived neurotrophic factor (BDNF) concentration as an indicator of polyneuropathy, which did not accurately reflect nerve damage and the results could not be generalized due to small number of patients. Only a limited number of molecular genetic studies in TiPN have been conducted, but none has discussed the association between genetic factors and TiPN in patients with CD. TiPN is thought to be a dose-limiting toxicity [12], and the therapeutic dose of thalidomide between chemotherapy and CD patients can differ by up to 10 times [13]. Thus, the existing findings are probably not fully representative of patients with CD.

The neurotoxic mechanism of thalidomide is still unclear [14]. Some hypotheses and treatment directions for TiPN have appeared. Thalidomide exhibits an antiangiogenic effect, which is considered to cause secondary ischemia and hypoxia of nerve fibers, which may lead to ischemia-related neuropathy. Vascular endothelial growth factor (VEGF) gene therapy in animals with TiPN led to obvious improvement in vascular recovery [15, 16]. In addition, neuronal susceptibility can be increased by the dysregulation of neurotrophic factors through immunomodulatory mechanisms [17]. Tonello *et al* analyzed dorsal root ganglion tissues from their animal experiments. Matrix metalloproteinase (MMP)9 monoclonal antibody significantly decreased oxidative stress and affected the expression of neuroinflammatory mediators, suggesting that MMP9 acts on peripheral nerve lesions. Thalidomide metabolites are considered to cause neuronal damage through reactive oxidative species causing damage to DNA [18]. BDNF is the only neurotrophic factor expressed during most peripheral damaged sensory neurons It has been proven to be an effective regulator of regeneration-related gene expression in the peripheral and central nervous systems [19, 20]. Additionally, Navia-Pelaez *et al* observed the induction of mouse neuralgia by ATP binding cassette transporter (ABC)A1/ATP binding cassette transporter (ABC)G1 knockdown, which prevented apolipoprotein A-I binding protein from reversing peripheral neuropathy allodynia [21]. Johnson *et al* reported that gene polymorphisms of

ABCA1 had a significant association with thalidomide-related neuropathy [8]. Interleukin (IL)-12 was upregulated in MPTP-intoxicated mice [22]. In the patients who received CTL019, IL-2 Level in those with neurotoxicity was higher than in those without neurotoxicity [23]. Zhang *et al* supported IL-12 cytokine profiles as indicators of neurotoxicity [24]. Few studies have demonstrated a relationship between cumulative dose and TiPN [25, 26]. The gene polymorphisms of transcriptional regulators, inflammatory cytokines and transporters may have an important impact on individual differences in TiPN. These genetic factors may adequately explain the neurotoxicity with clinical variables.

Therefore, an accurate model for identifying TiPN with comprehensive clinical and genetic variables is required in patients with CD. In recent years, powerful data mining and computing have encouraged a growing use of machine learning in the medical field, including diagnosis, treatment, prognostic data classification, and regression [27-29]. Tao *et al* generated a predictive model for clinical response in patients with rheumatoid arthritis using a multiomics approach and machine learning, with a predictive accuracy > 85% [30]. Mo *et al* developed a Tacrolimus (TAC) nephrotoxicity predictive model in nephrotic syndrome (NS) using machine learning algorithms with clinical and genetic variables, and 78% were accurately identified [31]. To date, there is no predictive model for CD patients generated by machine learning.

The aim of this study was to develop a sensitive and accurate TiPN predictive model based on clinical and genetic variables, which is beneficial for the treatment of thalidomide.

MATERIALS AND METHODS

Study population

A total of 164 CD patients diagnosed according to the criteria of Lennard-Jones were randomly recruited from the Sixth Affiliated Hospital of Sun Yat-sen University from January 2016 to June 2022. The patients treated with thalidomide at any course of their disease were considered eligible for the study. The inclusion criteria were: (1) diagnosis of CD; (2) CD Activity Index (CDAI) > 150 points, with endoscopically active lesions; and (3) refractory or intolerant to immunosuppressive drugs or biological agents which are used in current treatment. The exclusion criteria were: (1) fiber stenosis caused by gastrointestinal obstruction symptoms; (2) fistula, excluding anal fistula; (3) pregnancy or lactation; (4) fertility program during the study; (5) Less than eight weeks of biologic treatment after last IFX; (6) central or peripheral nervous disease; (7) abnormal liver and renal function; (8) heart dysfunction; (9) malignant tumor; and (10) active tuberculosis.

Ethical approval was obtained by the Ethics Committee of the Sixth Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China. This study was registered at the Clinical Trial Registry (Registration Number: NCT02956538). Written informed consent was obtained from all participants. Blood samples were collected from all recruited patients.

Assessment of neurotoxicity

For the neurophysiological assessment, the same methods were used throughout the whole process of observation to minimize any potential bias. For the clinical assessment, the National Cancer Institute Common Toxicity Criteria Sensory Scale (version 4.0, 2009) was used. Details of thalidomide administered to each patient were acquired directly from the medical records of the patients. Meanwhile, patients without blood samples or intact clinical data were excluded.

Clinical variables

In order to adjust the influence of clinical variables on TiPN, 18 types of baseline clinical variables were collected, including demographic data (weight, age, *etc.*), inflammatory indexes (C-reactive protein, erythrocyte sedimentation rate, *etc.*), hepatic function (alanine aminotransferase, aspartate aminotransferase, *etc.*), and routine blood examination. Table S1 shows the full names and local abbreviations of clinical characteristics.

Single nucleotide polymorphism selection and genotyping

We performed comprehensive detection of genetic variables (single nucleotide polymorphisms; SNPs), including genes related to thalidomide pharmacokinetic/pharmacodynamic pathways, metabolic enzymes, transcriptional regulators, nerve growth factors, inflammatory cytokines, *etc*. Details of the selection steps of these genes (SNPs) were as follows. (1) The physical position of these genes was obtained through the human Ensembl GRCh37 database (http://asia.ensembl.org/Homo_sapiens/Info/Index). In the VCF to PED Converter window (http://grch37.ensembl.org/Homo_sapiens/Tools/VcftoPed), positions of genes were entered, the Chinese Han population in Beijing was selected, and then, PED and info file for the SNPs of these genes were download. (2) Haploview software for entry criterion was set (minor allele frequency > 5%, r^2 <0.8, min genotype > 75%, and Hardy–Weinberg equilibrium > 0.05) to obtain the tag-SNP. Ultimately, 150 genetic variables met the above standards. DNA analysis was performed by collecting 5-mL peripheral blood samples. DNA extraction from whole blood was performed using Genomic Blood DNA Extraction Kit (DP304, Tiangen, Beijing, China). Nanodrop 2000C (Thermo Scientific, Fitchburg, WI, USA) was used for the detection of DNA concentration.

The published polymerase chain reaction-restriction fragment length polymorphism method was used for the detection of all SNPs [32, 33], and 150 SNPs were detected by Agena Bioscience MassARRAY (Agena Bioscience, San Diego, CA, USA).

Machine learning

Single-sample Kolmogorov–Smirnov tests were carried out to test the distribution of continuous variables. Data were shown as median (range) or mean ± SD, according to the data type.

Based on clinical and genetic variables, a predictive model for TiPN in CD patients was developed using a series of machine learning methods. We implement machine learning as a three-step process, data preprocessing, feature selection, and model generation and verification. To evaluate the performance of the model generation, a fivefold cross-verification was performed. The evaluation indicators used included confusion matrix, receiver operating characteristic (ROC) curve, precision recall (PR) curve, specificity, sensitivity, precision, accuracy and F1 core. Figure 1 shows the workflow for machine learning.

The *t*-test and nonparametric Mann–Whitney *U* test were used to analyze continuous variables while categorical variables were analyzed by c^2 test. Machine learning techniques were performed in Python 3.7.13. GraphPad Prism version 8 (GraphPad, San Diego, CA, USA) was used for graphic analysis. $P < 0.\ 05$ was considered to be a statistically significant difference.

Data preprocessing

Features with missing rates > 30% were removed after the collection of variables was finished. The missing value of a continuous variable was filled with an average value, and the classification variable was processed by removing the missing value. The continuous variables were uniform quantized by minimum–maximum normalization, and the categorical variables were represented using dummy variables. The range of each clinical data set is shown in Table S1.

Feature selection

Univariate analysis was used to evaluate the relationship between each variable and TiPN. SNPs with weak effects ($P > 0.1$) were eliminated. In order to reduce the model complexity, clinical variables were uniformly quantified into 11 categorical variables [0, 0.1, 0.2, ...,1], since the dataset included both categorical variables (e. g. gender, polymorphisms and outcomes) and continuous variables (e. g. drug dose, age and clinical tests). To improve robustness and accuracy of the prediction of our approach, the dataset was reiterated 1000 times, resulting in different test sets, which were used to select model hyperparameters. Five models were established to predict TiPN development, named extreme gradient boosting (XGBoost), gradient boosting decision tree (GBDT), extremely random tree (ET), random forest (RF) and logistic regression (LR) model. The paramount variables of all these models are ranked by information gain [34].

Model development and validation

One hundred and sixty-four patients were randomly divided into training (80%) and testing (20%) data sets [35, 36]. The Synthetic Minority Oversampling Technique （SMOTE ）algorithm was combined in the training set and the test set to deal with the data imbalance. The five predictive models were trained using k folder cross-validation (k = 5). Through implementation and comparison, we avoided merging plans and improved the generalization performance of these models.

XGboost, ET, GBDT, RF and LR algorithms were used to analyze the feature set and to generate neurotoxicity prediction models. To guarantee the robust stability of these computational models, the dataset was randomly divided 1000 times to obtain different training sets to test the hyperparameter for each group. Taking XGBoost as an example, the main hyperparameter included learning rate, maximum depth, estimators, and eta. The performance of the tested classification algorithm was evaluated and compared based on the area under the ROC (AUROC) curve, area under the PR (AUPR) curve, specificity, sensitivity (recall), precision, accuracy, and F1 score.

Bioinformatics analysis

The effects of expression quantitative trait loci (eQTLs) on the top-four gene expressions were examined with the Genotype-Tissue Expression (GTEx) database. https://www.gtexportal.org/home/.

RESULTS

Patient characteristics

A total of 164 patients with CD were collected in this study, including 132 men and 48 women. The average age of the patients was 34.3 ± 12.7 years. Median dose of thalidomide was 1.5 mg/kg/d (range 0.3–2.9). TiPN was observed in 59 patients (36%) during follow-up. Median duration of thalidomide treatment was 17.2 mo (range 1-60). We collected 18 baseline variables and 150 genetic variables of the patients. Table 1 shows the baseline characteristics of the patients. The genotype coincided with Hardy Weinberg equilibrium.

Feature selection

In the univariate analysis, variables that had a significant impact on TiPN included *IL-12* rs1353248 (P = 0.0004), *BDNF* rs2030324 (P = 0.001), *BDNF* rs6265 (P = 0.001), and dose (P = 0.002), *ABCA1* rs10991419 (P = 0.002). After removing genetic variables with low correlation (P > 0.1), 14 SNPs and 18 clinical variables were included in the following analysis (Tables 2 and S1).

Identification of appropriate biomarkers that distinguished the neurotoxic symptom group from the well-tolerated group was deemed important, which helped in exploring disease biomarkers and understanding pathogenesis. After the data transformation, models were generated using ET, GBDT, RF, XGBoost and LR, and the feature importance scores were used to rank all the features. The higher the information gain value, the more significant the variable became (Figure 2). By ranking each feature, the top five ranked features were *IL-12* rs1353248, dose (mg/d), *BDNF* rs6265, *BDNF* rs2030324 and *BDNF* rs11030104. *IL-12* rs1353248 was found to have the highest predictive variable importance for TiPN, followed by thalidomide daily dose, rs6265, rs2030324 and rs11030104. Figure 3 shows the four genetic variables and the occurrence of TiPN. Patients with *BDNF* rs2030324_AG, *BDNF* rs6265_CT, *BDNF* rs11030104_AG, and *IL-12* rs1353248_TT genotypes were more likely to have TiPN; patients with BDNF rs2030324_AG genotype had more neurotoxicity than patients with AA+GG; CT genotype carriers of *BDNF* rs6265 had higher neurotoxicity than CC+TT carriers; neurotoxicity in carriers of *BDNF* rs11030104_AG was more than in patients with AA+GG genotype; and patients with *IL-12* rs1353248_TT genotype had more neurotoxicity than those who carried CT+CC. Additionally, we noticed that there were three SNPs derived from the same gene, which suggested that BDNF played a significant role in neurotoxicity.

Functional consequences of the top four SNPs

The GTEx eQTL database was used to examine the functional consequences of the top four SNPs. The results suggested that four variants, rs1353248 (chr3_159905770, P = 8.52 ´ 10^{-4}), rs6265 (chr11_27658369, P = 1.07 ´ 10^{-4}), rs2030324 (chr11_27705368, P = 9.2 ´ 10^{-11}), rs11030104 (chr11_27662970, P = 2.76 ´ 10^{-5}) could affect IL-12 and BDNF gene expression on human nerve tibial tissue (Figure 4). The expression levels of the BDNF gene were reduced in rs6265CT and rs11030104AG, with a significant reduction in the rs6265CT genotype. Additionally, the expression levels of the IL-12 gene were significantly decreased in the rs1353248TT. The results showed a similar trend, indicating that these four loci may play an important biological role in peripheral neurotoxicity and have potential for the prediction of TiPN (Figure 3). Further investigation into their biological functions is warranted.

Comparison of five algorithms in the training set

Using k-fold cross-validation (k = 5) in the training set, we identified all possible parameter combinations identified by random grid search. The evaluation indicators used included confusion matrix, ROC curve, PR curve, specificity, sensitivity, precision, accuracy and F1 core. The average ROC curve, PR curve and 95% confidence interval (CI) are shown in Figure 5A and 5B. Here, the ROC curves of four models were > 0.90, the PR curve of the LR model was 0.874, and the remainder were > 0.97. In addition, these models exhibited different performances (Table 3). The ET model obtained the highest ROC curve (0.999) but had the lowest specificity rate of 0.471. The RF model obtained a high ROC curve (0.996) and a common precision (0.769) compared with XGboost and GBDT. The results indicated that XGboost and GBDT had superior performance, and the precision, specificity, accuracy and F1 scores were above 0.90, 0.82, 0.88 and 0.87, respectively.

To evaluate the importance of genetic variables or clinical features in five models, we used 14 SNPs, 18 clinical features, and three top five ranked features included in five models of the other three round workouts with similar analytic approach. The results indicated that by including 18 clinical features, ROC curves were 0.759–0.999, and PRC curves were 0.602–0.998 (Figure S1). The ET model showed the best sensitivity (1.0) but had the lowest precision (0.431) and specificity rate (0.262). The precision, specificity, accuracy, and F1 were low for LR and RF models (Table S3). Additionally, with only input of the 14 SNPs, the models achieved better performance than the clinical features. The ROC curve and PR curve of XGBoost, ET, GBDT, and RF models were > 0.90 (Figure S2); the precision, sensitivity, specificity, accuracy and F1 score of XGboost and GBDT were > 0.90 (Table S4). These results indicated that genetic variables were more important in predicting TiPN than clinical features.

When including the top five ranked features, the ROC curve for the four models was > 0.82, The LR model only achieved a low value (0.72). The PRC curve of the LR model was 0.578, and the remainders were > 0.72. The XGboost showed the highest PR curve (0.803, 95%CI 0.697–0.871) and accuracy (0.718). The GBDT obtained the highest ROC curve (0.88, 95%CI 0.804–0.956). However, F1 score was < 0.70 for all models (Figure S3, Table S5). The findings indicated that integration with genetic features and clinical data may be refining the performance of prediction models.

Validation of the five algorithms in the test set

Five algorithms in the test set were verified based on training results. The performance of five prediction models based on the 18 clinical features and 14 SNPs were: precision 0.625–0.8, sensitivity 0.667–1.0, specificity 0.667–0.889, accuracy 0.733–0.818, and F1 score 0.714–0.8 (Table 4). The average area of the ROC and PR curves was 0.741–0.907 and 0.718–0.864, respectively (Figure 6). The models generated by the XGBoost algorithm had the best overall predictive power and the highest specificity (0.857), accuracy (0.818) and PR curve (0.864, 95%CI 0.828–1.011); the ROC curve was (0.889, 95%CI 0.757–1.021), and the remaining values for precision, sensitivity and F1 score were > 0.75. The RF acquired the highest ROC curve (0.907, 95%CI 0.731–1.084), and the ET and GBDT achieved the best F1 score, but all three had the lowest specificity (0.667).

We validated only 14 genetic variables, 18 clinical variables and the top five ranked features in five models. All these models behaved poorly with the 18 clinical features (Table S6), while the sensitivity score was > 0.83 for four models (XGboost, ET, GBDT and RF). The performance of the overall value of the ROC curve (0.528–0.718) was higher than the PR curve (0.359–0.556) (Figure S4).

When only 14 SNPs were considered, the ROC curve of XGboost and RF were up to 0.802 and 0.907; the PRC curve was > 0.70 for the four models (XGboost, ET, GBDT and RF) (Figure S5), with accuracy, precision, sensitivity, specificity and F1 score above 0.73, 0.66, 0.66, 0.77 and 0.66, respectively (Table S7). The LR had the lowest ROC curve (0.722) and PR curve (0.601).

Considering only the top five ranked features, the ROC and PR curves were > 0.79 and > 0.73 for all these models, respectively (Figure S6). XGboost had the best overall predictive power and highest specificity (0.857), sensitivity (0.833), accuracy (0.848) and F1 score (0.8). The LR had the lowest accuracy (0.552), specificity (0.316) and precision (0.435) (Table S8).

DISCUSSION

Thalidomide leads to well-documented adverse effects in some patients, and drug discontinuation derived from neurotoxicity alone was up to 46% [2, 34]. However, it is still unclear to predict TiPN risk in Chinese people by combining genetic polymorphism and clinical factors.

So far, there have been few studies on the risk factors of TiPN in patients with CD. As far as we know, only Bramuzzo *et al* developed a model to identify genetic variables using LR to predict the occurrence of TiPN in children. They found that polymorphisms in *ICAM1* and *SERPINB2* were protective factors. However, genetic tests were used only in a few patients and the survey variables were not sufficient. The LR method also reduced the predictive performance of the model [26].

In comparison with traditional statistical methods, based on the previous studies, machine learning can generate models of higher predicted performance by handling more complex data, which may achieve higher accuracy and improved generalization [35,36]. With the rapid increase in artificial intelligence, machine learning methods are widely applied in the field of disease diagnosis and prediction [37-39]. Not only this, the method based on machine learning no longer requires strong assumptions about basic mechanisms such as image classification [38] and speech recognition [40], which have achieved cutting-edge predictive capabilities.

We developed a model for predicting TiPN in Chinese people using machine learning (XGBoost, ET, GBDT, RF and LR) based on genetic and clinical variables for the first time. As a result of the comprehensive evaluation, the model generated by the XGboost algorithm reached the optimum prediction ability, which could accurately distinguish 88.9% of patients (Table 4), this was consistent with the results of other relevant studies [41, 42]. By ranking each feature, rs1353248, rs6265, rs2030324 and rs11030104, and drug dose had the top five effects on TiPN. We showed that machine learning methods were superior to traditional statistical methods, and compared with two recently published studies on other diseases [43, 44], our XGboost model yielded higher ROC-AUC and PRC-AUC scores. Furthermore, the significant aspect of the XGboost model is that TiPN can be identified at a high probability early in the disease course and may significantly improve the treatment outcome.

We applied an additional analysis by inputting SNPs, clinical variables, and the top five ranked features investigate whether there was any impact on model performance. As a result, all the models performed well when including SNPs and the top five features, near the level of all features combining 18 clinical variables and 14 SNPs in the models. Taken together with the ROC and PR curves, these results showed that genetic variables had a more important role than clinical variables in predicting TiPN occurrence, while in combination with clinical data, they markedly improved the model, suggesting that integration with genetic features and clinical data refined the predictive models.

Among the four SNPs, we focused mainly on *BDNF* and *IL-12* according to the ranked results. So far, the relationship between SNPs of the *IL-12* gene and the risk of TiPN has not been studied. To our knowledge, only one genome-wide association study has reported that *IL-12* rs1353248 had strong relevance to celiac disease [45]. In our study, *IL-12* rs1353248 had distinct relevance to TiPN, indicating a new and genetically relevant connection between genetic determinants of IL-12 and TiPN risk.

BDNF plays a significant role in neuronal differentiation, survival, and synaptic plasticity. BDNF protein levels in fibromyalgia patients are significantly elevated. Similar to our research, Park *et al* found that *BDNF* rs11030104_GG had a protective effect against fibromyalgia compared with *BDNF* rs11030104_AG in a multicenter prospective study of the Korean population [46]. In addition, *BDNF* rs6265 SNP has been widely studied for its role in the regulation of neuronal survival, differentiation, and plasticity [47-50]. Xie *et al* revealed that rs2030324_CT and rs6265_AG were associated with amnestic mild cognitive impairment [51]. The effects of *BDNF* rs6265 polymorphism have been widely studied in animal models. In adult male mice, the *BDNF* Val66Met polymorphism impaired sports training-induced synaptic plasticity and beneficial behavior [52]. We reported similar findings, with *BDNF* rs6265_CT and *BDNF* rs2030324_AG being risk factors for TiPN. We suggest that heterozygotes for this gene impair neuronal activity, which is potentially involved in the perception of neuropathy symptoms. It can be speculated that mutation of the T allele of rs6265, G allele of rs2030324 and rs11030104 has a major impact on the expression levels of this gene according to the GTEx databases, and the mechanism of action merits further, in-depth investigation. In the Israeli population with lymphoma and myeloma, the severity and persistence of chemotherapy-induced peripheral neuropathy were markedly higher in the carriers of the Val/Val genotype than in patients with the Val/Met and Met/Met genotypes [48], which is different from our results. This difference may be due to the fact that the two research results are inherently different and difficult to compare directly. However, it suggests that the risk of neurotoxicity differs in different human populations according to the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/) and China Metabolic Analytics Project (http://www.mbiobank.com/). It is noteworthy that the mutation rate of rs6265 and rs11030104 in Asian populations is as high as 49%, while in European and American populations, the mutation rate of these loci does not exceed 22%. This indicates that a higher number of individuals in Asian populations may be susceptible to TiPN, and these two loci are of predictive significance in Asian populations.

In addition to SNP-based correlation analysis, we considered 18 clinical features that may be related to the development of TiPN. In the training set, five models (LR, RF, ET, GBDT and XGboost) were constructed to predict the TiPN risk including 14 SNPs and 18 clinical features. Considered collectively, the XGboost model obtained the optimum performance in all functional training and test sets. As a result of predicting the characteristic importance using the XGboost model, the dose was found to be the second most important factor. This finding was in line with previous reports [12, 53]. In addition, the prospective study on 135 patients with skin disease showed that the incidence rate of TiPN was 11% (4/35) with a dose of 25–50 mg/d, 29% (11/38) with 50–75 mg/d, and 48% (19/40) with 75–100 mg/d. The incidence of neurotoxicity was three times higher in patients taking higher doses than in those taking lower doses. Similar trends were found in this study, the incidence rate of TiPN was 18% (9/49) with 25–50 mg/d, 33% (14/43) with 50–75 mg/d, and 49% with > 75 mg/d.

There were some limitations to our study. First, the sample size was small because it was a single-center trial. To improve the generalization and robustness of our predictive model, we reiterated this random procedure 1000 times. Second, clinical measures like peripheral nerve injury severity, presence of current therapy, duration and cumulative dose were not specified by the initial protocol design. Third, a limited number of SNPs and genes were selected for examination because of cost constraints, and more SNPs remain to be found.

Interventions may be a key adjunct to reducing the incidence of TiPN in future treatment. This is the first study to use cutting-edge machine learning to establish and validate a TiPN predictive model using comprehensive genetic and clinical variables.

Genes encoding inflammatory cytokines, growth of nerve fibers, and enzymes involved in ubiquitination were also screened out. These SNPs are closely related to the occurrence and development of TiPN. The results suggest that SNPs are important to fully predict TiPN. Through the prediction of this model, physicians can assess the possibility of TiPN in CD patients, which contributes to the rational use, timely intervention after administration, and avoidance of peripheral nerve damage. These findings are important for the management of CD patients with thalidomide.

CONCLUSION

In this study, the XGBoost algorithm exhibited a high degree of accuracy in predicting TiPN by utilizing 18 clinical features and 14 genetic variables. Furthermore, it can identify high-risk patients through single nucleotide polymorphisms. This suggests that XGBoost may offer a feasible option for improving thalidomide efficacy in CD patients.

ARTICLE HIGHLIGHTS

Research background

Thalidomide-induced peripheral neuropathy (TiPN), a life-threatening condition in Crohn's disease, has a high incidence in Asia. However, there are no effective medical interventions for TiPN.

Research motivation

Can we develop a predictive model of TiPN combining genetic and clinical variables? Which variable affects TiPN more?

Research objectives

To establish an optimal model using clinical variables and genotypes to predict TiPN and improve the safety for the thalidomide treatment.

Research methods

A total of 164 patients diagnosed with Crohn's disease at the Sixth Affiliated Hospital of the Sun Yat-Sen University were included in this study. Peripheral blood was collected from the patients to detect the genotypes at School of Pharmaceutical Sciences, Sun Yat-Sen University. The X^2 method or Single-sample Kolmogorov–Smirnov test was used to determine the association of TiPN with 18 clinical features and 150 genetic variables. Five predictive models were established and evaluated by the confusion matrix receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), specificity, sensitivity (recall rate), precision, accuracy, and F1 score.

Research results

TiPN was observed in 59 individuals. Among the five models, XGBoost algorithm exhibited the best predictive performance with the highest specificity (0.857), accuracy (0.818), AUPRC (0.86) and AUROC (0.89) after evaluation. The top-ranking five risk variables associated with TiPN were interleukin-12 rs1353248 [*P* = 0.0004, odds ratio (OR) 8.983, 95% confidence interval (CI) 2.497–30.90), dose (mg/d, *P* = 0.002), brain-derived neurotrophic factor (BDNF) rs2030324 (*P* = 0.001, OR 3.164, 95%CI 1.561–6.434), BDNF rs6265 (*P* = 0.001, OR 3.150, 95%CI 1.546–6.073) and BDNF rs11030104 (*P* = 0.001, OR 3.091, 95%CI 1.525–5.960).

Research conclusions

The XGBoost algorithm accurately predicts TiPN using 18 clinical features and 14 genetic variables. It is able to identify high-risk patients using single nucleotide polymorphisms.

Research perspectives

Applying the machine learning to adjust thalidomide therapies based on these specific genotypes is recommended before the thalidomide treatment.

ACKNOWLEDGEMENTS

8

We thank the physicians and nurses from department of gastroenterology for their cooperation in this study.

84842_Auto_Edited.docx

ORIGINALITY REPORT

3%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet	56 words — 1%
2	www.wjgnet.com Internet	54 words — 1%
3	Yang Li, Maryam B. Lustberg, Shuiying Hu. "Emerging Pharmacological and Non-Pharmacological Therapeutics for Prevention and Treatment of Chemotherapy-Induced Peripheral Neuropathy", Cancers, 2021 Crossref	21 words — < 1%
4	www.jlr.org Internet	15 words — < 1%
5	id.nii.ac.jp Internet	14 words — < 1%
6	translational-medicine.biomedcentral.com Internet	14 words — < 1%
7	doctorpenguin.com Internet	13 words — < 1%
8	Xiaolan Mo, Xiujuan Chen, Huasong Zeng, Wei Zheng et al. "Tacrolimus in the treatment of childhood nephrotic syndrome: machine learning detects novel	12 words — < 1%

biomarkers and predicts efficacy", Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 2022
Crossref

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 12 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS