

Research Article

Predictive Model for Nonadherence to Biologic Therapy in Ulcerative Colitis Patients Based on Machine Learning

Xian Huang , Yan Zhang , Yanru Wang , Huamei Lin , Yuzhu Yang* 

Department of Anorectal Surgery VI, Shenzhen TCM Anorectal Hospital (Futian), Shenzhen, China

Abstract

Objective: To develop a clinical prediction model for nonadherence to biologic therapy in patients with ulcerative colitis (UC) using machine learning techniques, and to assess its predictive accuracy to guide clinical interventions. **Methods:** A total of 221 UC patients who initiated biologic therapy between December 2023 and February 2025 at our hospital were included in this study. Data on 15 variables, such as age, sex, disease duration, and other clinical factors, were collected. Medication adherence was measured using the proportion of days covered (PDC), with a PDC >80% considered indicative of good adherence. The Support Vector Machine (SVM) and Boruta algorithms were employed to identify key predictors. A multivariate logistic regression model was developed using the intersection of factors identified by both algorithms. Model performance was evaluated using the C-index, ROC curve, calibration curve, decision curve analysis, and validation with the K-nearest neighbor (KNN) algorithm. **Results:** Of the 221 patients, 82 (37.1%) were categorized into the nonadherence group. Following factor selection, insurance status, depression level, education level, disease activity, and age were identified as significant predictors of nonadherence. The model demonstrated a C-index of 0.779, and the ROC curve showed an area under the curve (AUC) of 0.78. Calibration curve analysis revealed good model consistency, and KNN validation yielded high precision with an AUC of 0.9978 and a PR AUC of 0.9963. **Conclusion:** The developed prediction model for medication nonadherence in UC patients demonstrates robust predictive and calibration capabilities. This model may aid healthcare professionals in identifying high-risk patients and supporting timely clinical interventions.

Keywords

Ulcerative Colitis, Machine Learning, Clinical Prediction Model, Boruta, SVM-REF

1. Introduction

Ulcerative colitis (UC) is a chronic condition characterized by diffuse inflammation of the rectal and colonic mucosa. In 95% of cases, UC affects the rectum, and it may extend to the more proximal parts of the colon. The hallmark symptom of UC is bloody diarrhea. The incidence of UC varies widely between 0.5 and 31.5 per 100,000 people, depending on the population and region, with rates in Asia ranging from 5.3 to

63.6 per 100,000 individuals [1, 2]. The peak onset of UC occurs between the ages of 30 and 40 years [3]. Due to its unclear etiology, high recurrence risk, poor prognosis, impact on daily work, and diminished quality of life, UC presents a significant clinical challenge.

Nonadherence to medication refers to the behavior of discontinuing treatment within the prescribed period [4]. The

*Corresponding author: szyyz0755@163.com (Yuzhu Yang)

Received: 17 April 2025; **Accepted:** 27 April 2025; **Published:** 22 May 2025



Copyright: © The Author(s), 2025. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

consequences of nonadherence are numerous, including disease relapse, worsening symptoms, increased morbidity, and even mortality. Multiple factors influence medication nonadherence, such as socioeconomic status, medication dosage, side effects, income, and educational level. Given the multitude of factors influencing nonadherence, the development of accurate predictive tools and early interventions may be one of the most effective methods to improve adherence [5].

This study aims to develop an effective and simple predictive tool to assess the risk of nonadherence to biologic therapy in UC patients using machine learning techniques.

2. Materials and Methods

2.1. Study Population

A total of 221 UC patients who started biologic therapy between December 2023 and February 2025 at our hospital were included in the study. All enrolled patients were newly initiated on medication therapy. Inclusion criteria: (1) diagnosis according to the American College of Gastroenterology (ACG) 2019 guidelines for UC [6]; (2) basic literacy skills and effective communication; (3) signed informed consent. Exclusion criteria: (1) severe psychiatric disorders or dementia; (2) comorbid malignant tumors; (3) history of allergy to the prescribed medication. This clinical study was approved by the Hospital Ethics Committee (Approval No. GCYY202301005).

2.2. Methods

Data collected included patient age, disease duration, sex, smoking and drinking habits, insurance status, monthly income, education level, disease activity, medication frequency, anxiety and depression scores, physician-patient trust, and family medical history. Physician-patient trust was assessed using the Chinese version of the Wake Forest Physician Trust Scale [7], while anxiety and depression were measured using the Hamilton Anxiety and Depression Scales [8, 9]. Disease activity was assessed by erythrocyte sedimentation rate (ESR). After the first treatment, follow-up was conducted by phone or WeChat at 1 week, 1 month, 2 months, and 6 months, to assess adverse drug reactions (e.g., skin rashes, liver or kidney dysfunction, injection site reactions, and infections). A total of 15 variables were considered as potential predictors.

Medication adherence was evaluated using the proportion of days covered (PDC) [10]. A PDC greater than 80% was considered good adherence [11]. Follow-up visits were conducted at 1, 2, and 6 months to monitor adherence, with PDC calculated from prescription records.

2.3. Statistical Analysis

Data were input and cross-checked using Microsoft Excel, with any missing values exceeding 50% removed. KNN im-

putation was applied to continuous variables (age, disease duration) and random forest imputation was used for categorical variables (insurance status, education level). This hybrid approach aimed to preserve data distribution patterns while minimizing bias..

Factor transformation was performed before analysis. Age was categorized into <50 years (0) and ≥ 50 years (1); sex was coded as male (0) and female (1); smoking and drinking history was coded as absent (0) or present (1); insurance status was coded as 0 (no insurance) and 1 (insured); income was divided into <5000, 5000-10000, and >10000 RMB, and coded as 0, 1, and 2, respectively; education level was categorized as \leq middle school (0), high school (1), and college or higher (2). Disease activity was categorized based on ESR, with values <20 mm/H (0), 20-60 mm/H (1), 60-100 mm/H (2), and >100 mm/H (3). Medication side effects were coded as absent (0) or present (1), and medication frequency was coded as 1 (once per month), 2 (twice per month), or 3 (more than twice per month). Anxiety and depression scores were categorized based on the Hamilton scales. Family medical history was coded as absent (0) or present (1).

SVM and Boruta algorithms were used for feature selection. SVM analysis was performed using the e1071 package in R, with a linear kernel and a penalty factor (C) set to 0.1. The Boruta algorithm was run with a maximum of 60 iterations and a doTrace parameter of 2. The final predictive model was built using multivariate logistic regression and visualized using a nomogram.

2.4. Model Evaluation

Model performance was assessed using the C-index, ROC curve, calibration curve, and decision curve analysis. KNN algorithm validation was performed to verify model accuracy, with performance metrics including ROC AUC, PR AUC, confusion matrix, accuracy, sensitivity, specificity, F1 score and 5-fold cross-validation.

3. Results

3.1. Feature Selection and Model Construction

A total of 221 patients were included, with 139 in the adherence group and 82 in the nonadherence group (37.1%). The Boruta algorithm identified five factors: education level, insurance status, depression level, disease activity, and age. The SVM algorithm identified seven factors, including smoking history and medication frequency. The intersection of factors identified by both methods led to the inclusion of insurance status, depression level, education level, disease activity, and age in the predictive model (Figure 1). A multivariate logistic regression model was constructed using these five factors, and the results were visualized using a nomogram (Figure 2).

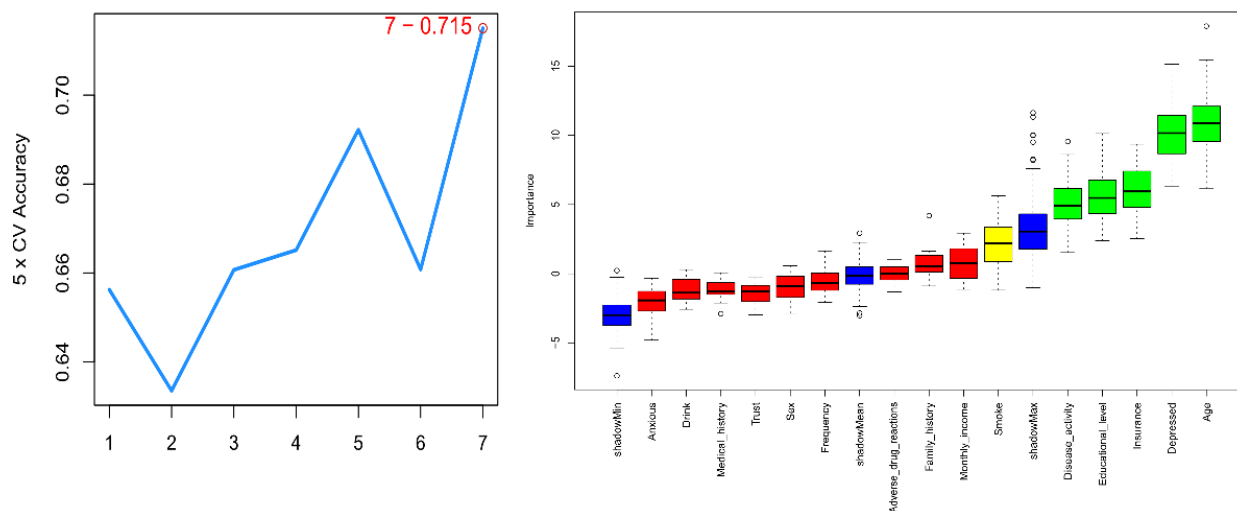


Figure 1. Key Gene Selection (A. SVM-REF B. Boruta).

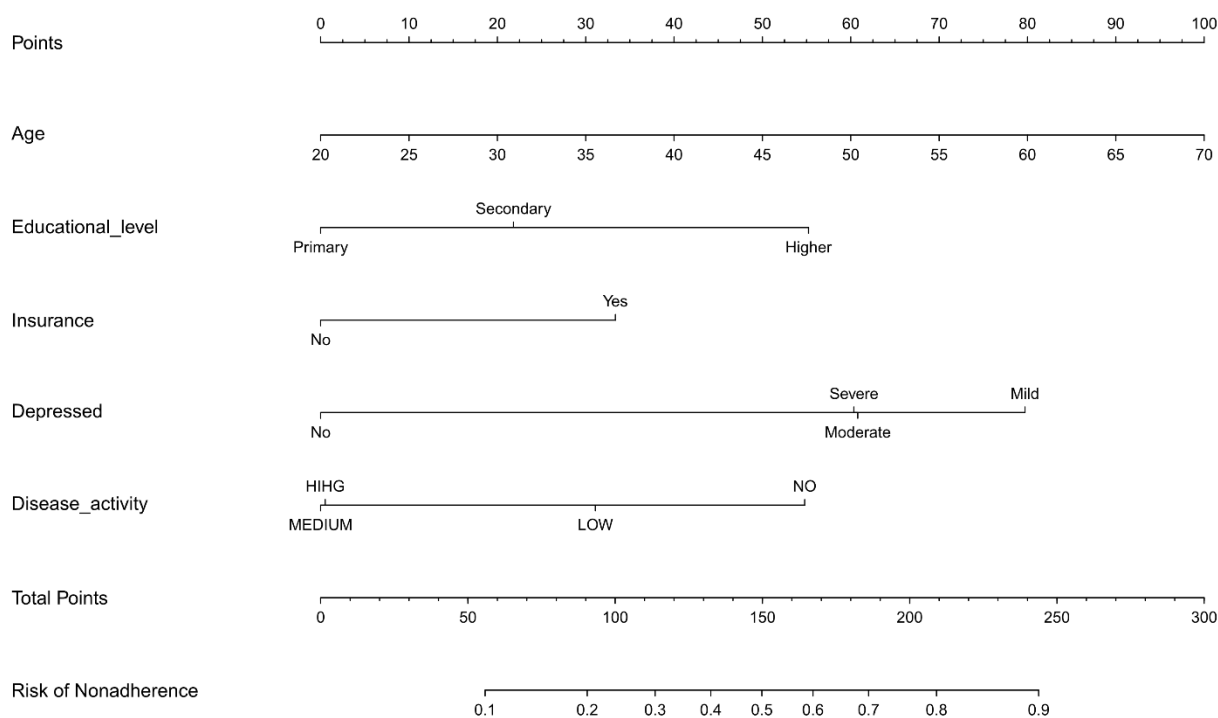


Figure 2. Nomogram of medication Nonadherence.

3.2. Model Performance Evaluation

The C-index of the model was 0.779, indicating good discrimination. The calibration curve showed excellent con-

sistency (Figure 3A). The ROC curve revealed an AUC of 0.78, suggesting high accuracy (Figure 3B). The decision curve analysis indicated that approximately 90% of patients would benefit without compromising other patients' outcomes (Figure 4).

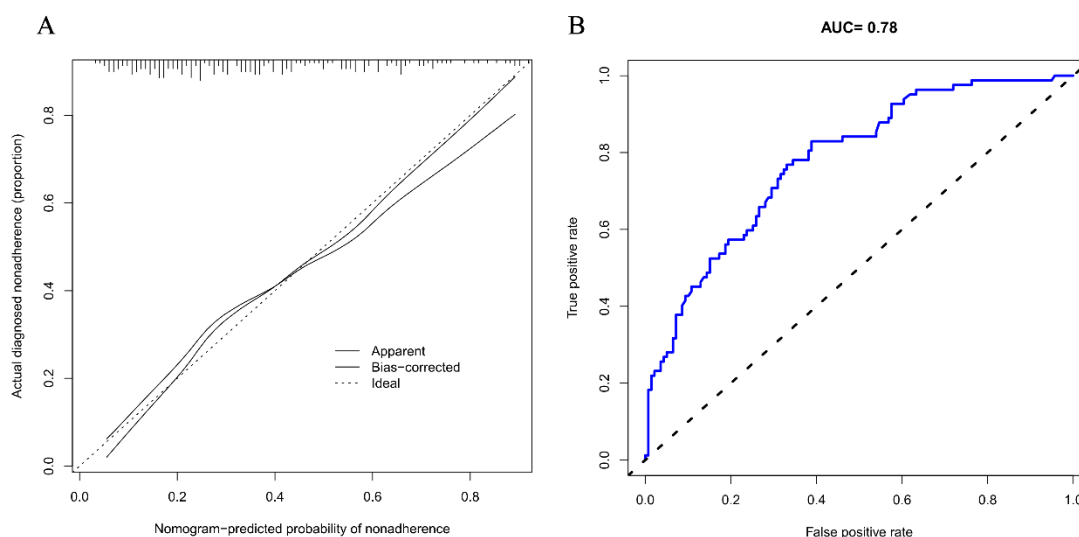


Figure 3. Clinical decision curve and ROC curve of the model.

A: The dotted line of the diagonal represents the ideal model, and the solid line represents the actual performance of the nomogram. The closer the solid line is to the dotted line of the diagonal, the stronger the prediction ability of the model. B: The ROC curve shows that the area under the curve (AUC) is 0.78.

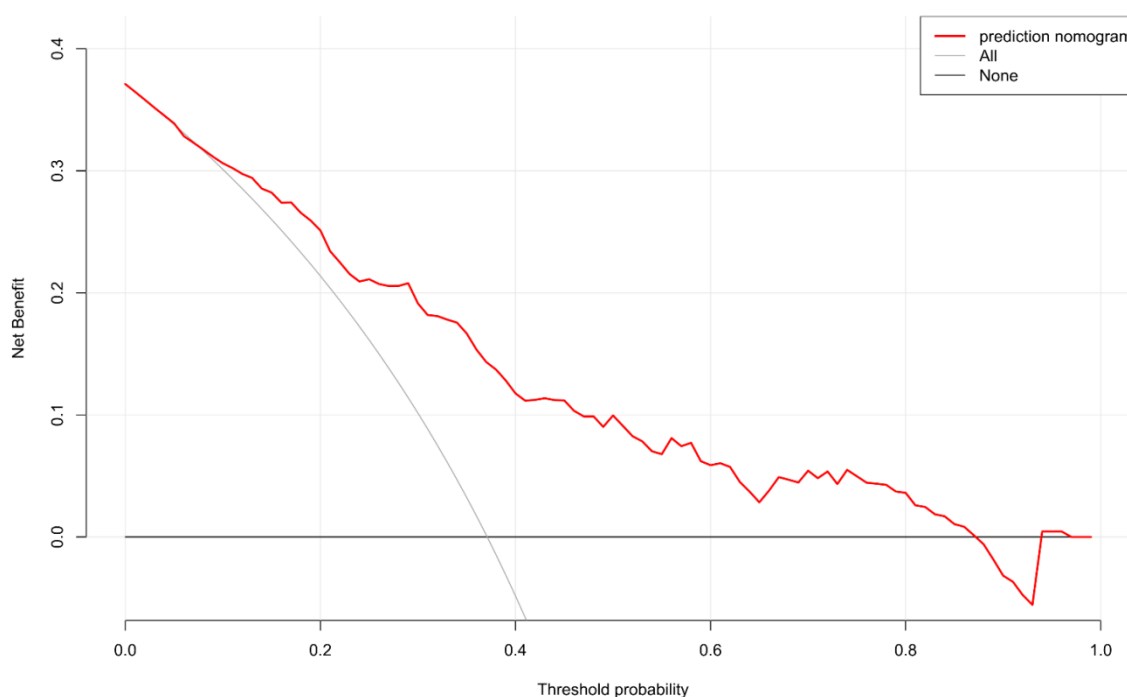


Figure 4. Clinical decision curve.

3.3. KNN Algorithm Validation

The KNN algorithm validation showed an ROC AUC of 0.9978 and a PR AUC of 0.9963, demonstrating high preci-

sion (Figure 5). The confusion matrix results showed a true positive rate (TPR) of 0.987, a true negative rate (TNR) of 0.968, accuracy of 0.968, and an F1 score of 0.958 (Figure 6). Five-fold cross-validation confirmed the model's robustness (ROC > 0.9; Figure 7).

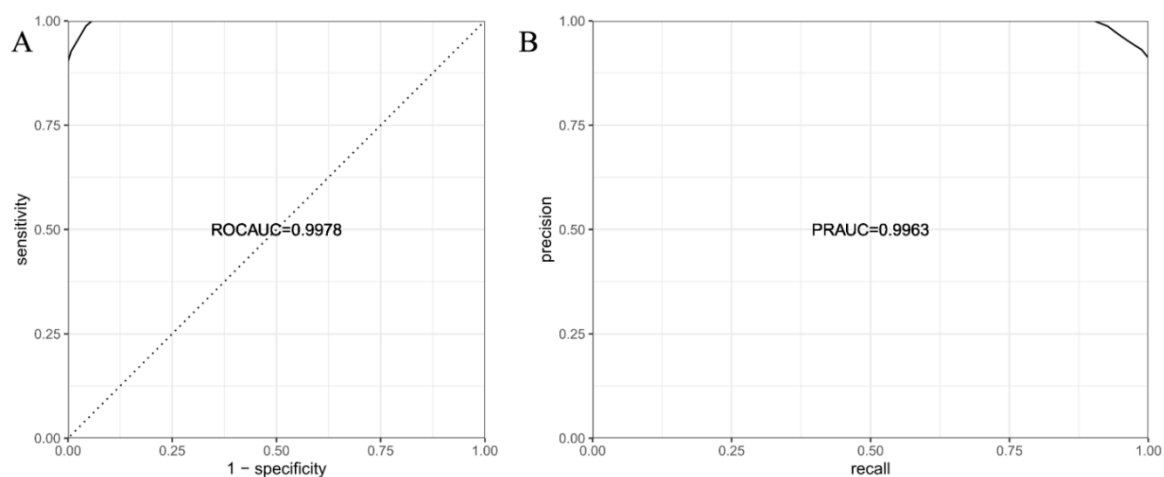


Figure 5. Verifying the prediction ability of the model by the KNN algorithm.

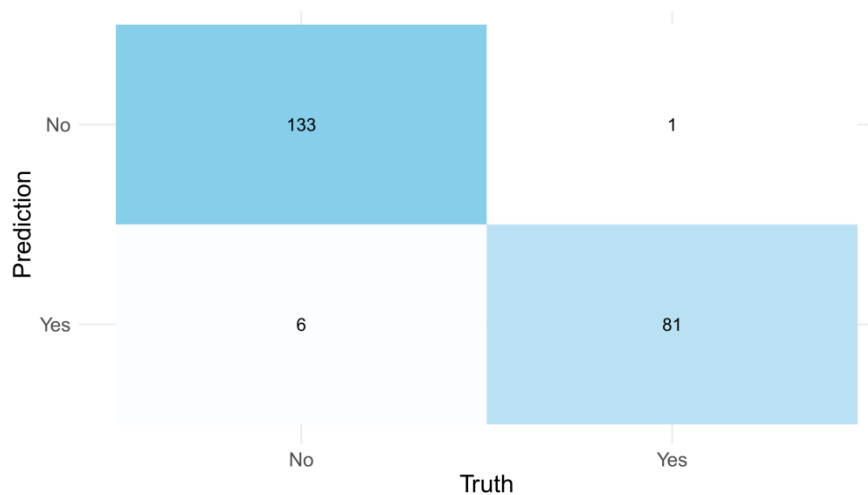


Figure 6. Confusion matrix.

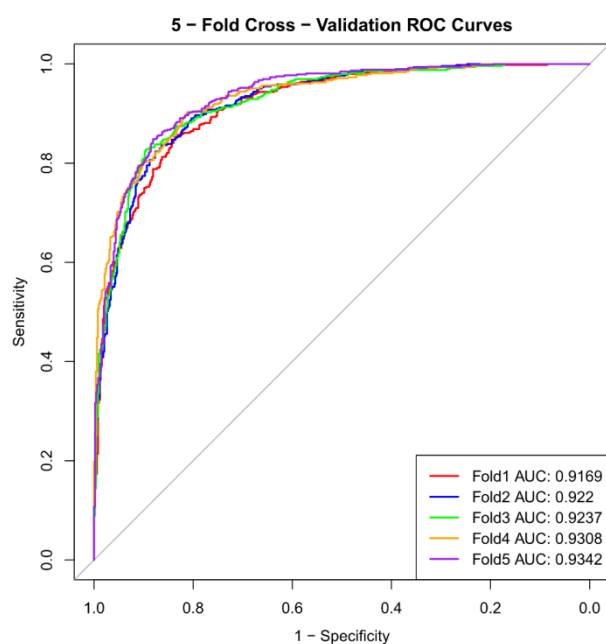


Figure 7. Five-fold cross-validation.

4. Discussion

Machine learning has been increasingly applied in the prediction of chronic diseases, identification of disease risk factors, and forecasting treatment outcomes. Several research teams have developed risk prediction models for inflammatory bowel disease (IBD) using gene expression datasets and machine learning (ML). For instance, in a cross-sectional study by Isakov et al., which involved 180 patients with Crohn's disease (CD), 149 patients with ulcerative colitis (UC), and 90 healthy controls, random forest (RF) and support vector machine (SVM) algorithms were used to classify a set of 16,390 genes from microarray and RNA-seq datasets. Their IBD risk prediction model demonstrated an area under the curve (AUC) of 0.829, with sensitivity, specificity, and accuracy values of 0.577, 0.880, and 0.808, respectively [12]. This study, in contrast, employed various machine learning methods to identify multiple factors and construct a clinical prediction model specifically for non-adherence to biologic therapy in UC patients, with the aim of precisely identifying those at risk of poor adherence and enabling timely interventions. Validation of this model showed good predictive and calibration capabilities, as evidenced by the C-index, ROC curve, precision-recall (PR) curve, and other performance measures..

The developed non-adherence prediction model can assist clinicians in identifying UC patients at high risk for non-adherence. This study revealed that 37.1% of UC patients did not adhere to their prescribed medication regimen. Key risk factors for non-adherence were found to include health insurance status, depression levels, education level, disease activity, and age. Jackson et al. demonstrated that psychological distress, such as depression, anxiety, psychiatric diagnoses, or chronic perceived stress, negatively affects medication adherence in UC patients [4]. A study by Xu et al. indicated that in IBD patients, medication frequency significantly impacted the Morisky score, with patients taking medications four times a day scoring lower than those with once-daily dosing. More frequent dosing was thus associated with lower adherence [13]. Age was also found to be a significant factor in adherence, with an Italian study confirming that non-adherence was notably higher in patients under 40 years old ($P=0.041$) [14]. Economic factors, such as the cost of treatment and lack of medical insurance, also play a crucial role in medication adherence. Financial pressures and medication costs often compel patients to make difficult decisions regarding the timing and frequency of their medication. Ediger et al. identified medical costs as the most common barrier to medication adherence in a study of 326 newly diagnosed Canadian IBD patients, with 25% of respondents citing financial constraints as a significant obstacle to continuing treatment [15]. Kane further noted that as IBD progresses, adherence rates tend to decrease, and non-adherence in hospitalized and outpatient IBD patients is closely linked to high

medical costs [16].

Previous studies have also demonstrated the impact of medication adherence on disease activity in UC patients. In a retrospective study, Robinson et al. reported that the risk of disease relapse was significantly higher among non-adherent patients compared to those who adhered to treatment (odds ratio [OR] = 1.44, 95% confidence interval [CI] = 1.08–1.94; $P = 0.014$) [17]. Additionally, a study by Mood et al. found that non-adherence or discontinuation of sulfasalazine treatment increased the risk of colorectal cancer. Among 152 patients, five of the 16 non-adherent individuals developed colorectal cancer [18]. This finding was further supported by Eaden et al., who conducted a case-control study of 102 UC colorectal cancer cases. They found that patients receiving routine mesalazine treatment (≥ 1.2 g/day) had an 81% lower risk of developing colorectal cancer compared to untreated patients ($P = 0.006$) [19].

5. Conclusion

In conclusion, the predictive model developed in this study for nonadherence to biologic therapy in ulcerative colitis (UC) patients demonstrates robust predictive accuracy and calibration capabilities, providing a valuable tool for clinicians to identify high-risk patients and implement early interventions. Machine learning was prioritized over traditional regression for feature selection due to its ability to capture non-linear relationships and complex interactions among variables (e.g., synergistic effects between depression severity and socioeconomic factors), which are often overlooked by linear regression frameworks. However, several limitations should be acknowledged. First, the model was derived from a single-center cohort in China, which may limit its external generalizability due to region-specific healthcare policies and socioeconomic factors. Multicenter and cross-cultural validations are essential to confirm its applicability in diverse populations. Second, the exclusion of comorbid conditions and dynamic treatment response indicators might have omitted potential predictors. Additionally, while the model integrates machine learning for feature selection, its logistic regression framework may oversimplify complex interactions among variables.

Future research should prioritize the integration of real-time adherence monitoring technologies, such as electronic pill monitors or mobile health platforms, to dynamically track behavioral patterns and enhance model precision. Additionally, international multicenter validation is critical to evaluate the model's robustness across diverse healthcare systems and socioeconomic contexts. Further exploration of hybrid methodologies that synergize machine learning with causal inference frameworks could refine the balance between predictive power and clinical interpretability. Through iterative refinement, this model has the potential to evolve into a versatile decision-support tool for personalized ulcerative colitis management.

Abbreviations

| | |
|-----|-----------------------------------|
| UC | Ulcerative Colitis |
| ESR | Erythrocyte Sedimentation Rate |
| ROC | Receiver Operating Characteristic |
| SVM | Support Vector Machine |
| KNN | K - Nearest Neighbor |

Author Contributions

Conceptualisation: Huamei Lin. Data curation: Xian Huang. Data analysis: Xian Huang, Yuzhu Yang, Yan Zhang. Writing-original draft: Xian Huang, Yanru Wang. Approval of final manuscript: all authors.

Funding

The present study was funded by the Shenzhen Futian District Health and Public Welfare Research Project (grant no. FTWS2023010).

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Kobayashi T, Siegmund B, Le Berre C, et al. Ulcerative colitis. *Nat Rev Dis Primers*. 2020; 6: 74. <https://doi.org/10.1038/s41572-020-0205-x>
- [2] Asakura K, Nishiwaki Y, Inoue N, et al. Prevalence of ulcerative colitis and Crohn's disease in Japan. *J Gastroenterol* 2009; 44: 659-665. <https://doi.org/10.1007/s00535-009-0057-3>
- [3] Fumery M, Singh S, Dulai PS, et al. Natural history of adult ulcerative colitis in population-based cohorts: a systematic review. *Clin Gastroenterol Hepatol*. 2018; 16(343-356): e343. <https://doi.org/10.1016/j.cgh.2017.06.016>
- [4] Jackson CA, Clatworthy J, Robinson A, et al. Factors associated with non-adherence to oral medication for inflammatory bowel disease: a systematic review. *Am J Gastroenterol* 2010; 105(3): 525-539. <https://doi.org/10.1038/ajg.2009.685>
- [5] Vangeli E, Bakhshi S, Baker A et al. A systematic review of factors associated with non-adherence to treatment for immune-mediated inflammatory diseases. *Adv Ther*. 2015; 32(11): 983-1028. <https://doi.org/10.1007/s12325-015-0256-7>
- [6] Maaser C, Sturm A, Vavricka SR, et al. ECCO-ESGAR guideline for diagnostic assessment in IBD part 1: initial diagnosis, monitoring of known IBD, detection of complications *J. J Crohns Colitis*, 2019, 13(2): 144-164. <https://doi.org/10.1093/ecco-jcc/jjy113>
- [7] Dong Enhong, Bao Yong. Reliability and Validity of the Chinese Revised Version of the Wake Forest Physician Trust Scale. *Chinese Journal of Mental Health*, 2012, 26(3): 171-175.
- [8] Hamilton M. The assessment of anxiety states by rating. *Br J Med Psychol*, 1959, 32(1): 50-55. <https://doi.org/10.1111/j.2044-8341.1959.tb00467.x>
- [9] Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960; 23: 56-62. <https://doi.org/10.1136/jnnp.23.1.56>
- [10] Chu LH, Kawatkar AA, Gabriel SE. Medication adherence and attrition to biologic treatment in rheumatoid arthritis patients. *Clin Ther*. 2015, 37(3): 660-666.e8. <https://doi.org/10.1016/j.clinthera.2014.10.022>
- [11] Choudhry NK, Shrank WH, Levin RL, et al. Measuring concurrent adherence to multiple related medications. *Am J Manag Care*. 2009, 15(7): 457-464. <https://doi.org/10.1345/ajmc.15.7.457>
- [12] Yuan H, Ali M S, Brouwer E S, et al. Real - World Evidence: What It Is and What It Can Tell Us According to the International Society for Pharmacoepidemiology (ISPE) Comparative Effectiveness Research (CER) Special Interest Group (SIG) [J]. *Clinical Pharmacology & Therapeutics*, 2018, 14(2). <https://doi.org/10.1002/cpt.1086>
- [13] Xu F, Tang J, Zhu Z, et al. Medication Adherence and Its Influencing Factors Among Inflammatory Bowel Disease Patients in China. *Int J Gen Med*. 2022; 15: 4141-4149. <https://doi.org/10.2147/IJGM.S359082>
- [14] D'Incà R, Bertomoro P, Mazzocco K, et al. Risk factors for non-adherence to medication in inflammatory bowel disease patients. *Aliment Pharmacol Ther*. 2008; 27(2): 166-172. <https://doi.org/10.1111/j.1365-2036.2007.03555.x>
- [15] Ediger JP, Walker JR, Graff L, et al. Predictors of medication adherence in inflammatory bowel disease. *Am J Gastroenterol* 2007; 102: 1417-1426. <https://doi.org/10.1111/j.1572-0241.2007.01212.x>
- [16] Kane SV. Systematic review: adherence issues in the treatment of ulcerative colitis. *Aliment Pharmacol Ther* 2006; 23: 577-585. <https://doi.org/10.1111/j.1365-2036.2006.02809.x>
- [17] Robinson A, Hankins M, Wiseman G, et al. Maintaining stable symptom control in inflammatory bowel disease: a retrospective analysis of adherence, medication switches and the risk of relapse. *Aliment Pharmacol Ther*. 2013; 38: 531-538. <https://doi.org/10.1111/apt.12396>
- [18] Moody GA, Jayanthi V, Probert CS, et al. Long-term therapy with sulphasalazine protects against colorectal cancer in ulcerative colitis: a retrospective study of colorectal cancer risk and compliance with treatment in Leicestershire. *Eur J Gastroenterol Hepatol*. 1996; 8(12): 1179-1183. <https://doi.org/10.1097/00042737-199612000-00009>
- [19] Eaden J, Abrams K, Ekbom A, et al. Colorectal cancer prevention in ulcerative colitis: a case-control study. *Aliment Pharmacol Ther*. 2000; 14(2): 145-153. <https://doi.org/10.1046/j.1365-2036.2000.00698.x>