Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and External Validation of a Clinically Interpretable Prediction Model for Differentiating Crohn's Disease from Intestinal Tuberculosis

Ruibin Xu[1]
Shenzhen Third People's Hospital
Xiaoyin Zhang[1]
2350959447@qq.com

Shenzhen Third People's Hospital
Chuyao Zhang[1]
Shenzhen Third People's Hospital
Dayong Sun[2]
Shenzhen Second People's Hospital
Weitao Wu[2]
Shenzhen Second People's Hospital

Research Article

Keywords: Crohn's disease, intestinal tuberculosis, differential diagnosis, clinical prediction model

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11330365/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Crohn's disease (CD) and intestinal tuberculosis (ITB) have overlapping clinical, endoscopic, radiologic, and histopathologic features, complicating diagnosis when confirmatory microbiologic or characteristic histopathologic evidence is unavailable. We developed a practical prediction model using routinely available diagnostic information.

Methods

A multivariable logistic regression model was developed in 132 complete cases (82 with CD and 50 with ITB) and evaluated using a fixed formula and cutoff in a same-center temporal validation cohort (n = 117) and an independent external validation cohort (n = 65). Discrimination, calibration, and classification performance were assessed.

Results

The final model included age at symptom onset, interferon-gamma release assay, and the numbers of endoscopic features supporting CD and ITB. The AUC was 0.970 (95% CI, 0.947–0.994) in development, 0.9994 (95% CI, 0.9980-1.0000) in temporal validation, and 0.9995 (95% CI, 0.9981-1.0000) in external validation. Accuracy was 97.4% and 95.4% in the two validation cohorts, respectively. The model generally underestimated ITB probability.

Conclusions

The concise model combining age at symptom onset, IGRA, and structured endoscopic features showed promising discrimination and classification performance in the validation cohorts and may assist CD-ITB differentiation. Prospective multicenter studies should evaluate calibration, clinical utility, and generalizability.

Key Message

• **What is already known?** Crohn’s disease (CD) and intestinal tuberculosis (ITB) share overlapping clinical and endoscopic features, making accurate differentiation challenging in routine practice.

• **What is new here?** A concise model combining age, IGRA, and structured endoscopic features showed high discrimination in same-center temporal and independent external validation.

• **How can this study help patient care?** This model may provide a practical adjunct to support the differentiation of CD from ITB and facilitate more informed diagnostic decision-making.

Introduction

Crohn's disease (CD) and intestinal tuberculosis (ITB) may both present with chronic granulomatous intestinal inflammation and show substantial overlap in clinical symptoms and endoscopic, radiologic, and histopathologic findings. The burden of inflammatory bowel disease, particularly CD, continues to increase in Asia, while Asia remains a region with substantial epidemiological overlap between inflammatory bowel disease and tuberculosis, resulting in persistent epidemiologic overlap between the two diseases [1–5]. A long-term Korean cohort reported that 17.9% of patients with CD had initially been diagnosed with ITB and that 10.8% of patients with ITB had initially been diagnosed with CD [6]. Because the treatment pathways differ substantially, misdiagnosis may result in unnecessary antituberculosis therapy, delayed appropriate treatment for CD, or tuberculosis progression after corticosteroids, immunosuppressants, or biologic agents are administered to patients with unrecognized ITB [2, 6]. Reliable differentiation before the initiation of disease-specific treatment is therefore clinically important.

Differentiation between CD and ITB generally requires integration of evidence from multiple sources. The eight endoscopic features proposed by Lee et al. provide a basis for structured assessment [7], but the discriminatory value of endoscopic features alone may be limited in patients with early or mild disease and when lesions extend beyond the reach of colonoscopy. Routine markers such as complete blood count parameters, C-reactive protein, and erythrocyte sedimentation rate reflect systemic inflammation, but abnormalities occur in both diseases and lack specificity. Positive tuberculosis-related immunologic or microbiologic test results increase the likelihood of ITB; however, paucibacillary disease, sampling error, and the inability of some immunologic tests to distinguish latent infection from active disease mean that negative results do not exclude ITB [1–2, 8–9]. Imaging and histopathology can provide extramural and histologic clues, respectively, but their diagnostic performance remains limited by examination quality, reader expertise, and sensitivity. No single source of evidence is therefore sufficient for reliable differentiation, and findings from multiple domains must be interpreted together. Previous research has progressed from individual features and endoscopic scores to prediction models that integrate multidimensional information [10–23]. Candidate biomarkers, including serum proteins, anti-Saccharomyces cerevisiae antibodies (ASCA), immune-cell subsets, and microRNAs, have also been investigated [24–27].

Existing models support the value of multidimensional integration, but they differ considerably in the number of predictors, model structure, and validation strategy. A multicenter Asian validation study further showed that the high AUC and classification performance observed at the original cutoff in a development cohort may not be fully reproduced at a new center [17]. We therefore aimed to integrate diagnostic indicators into a parsimonious, easy-to-calculate model for differentiating CD from ITB and to evaluate the model through same-center temporal validation and independent external validation. We also compared the model with IGRA and the Lee endoscopic score, analyzed misclassified cases, and

performed independent repeat interpretation of endoscopic images to further assess the model's classification characteristics and practical implementation considerations.

Materials and Methods

Study Design and Participants

The study comprised retrospective model development, bootstrap resampling to assess coefficient stability, same-center temporal validation, and independent external validation. Clinical data were retrospectively collected from patients with a final diagnosis of CD or ITB at Shenzhen Third People's Hospital between January 1, 2018, and December 31, 2024. The same-center temporal validation cohort continuously included patients from the same hospital between January 1, 2025, and June 30, 2026.The independent external validation cohort included patients with CD or ITB identified at Shenzhen Second People's Hospital between January 1, 2022, and June 30, 2026. All data were coded and de-identified after extraction.

The size of the development cohort was determined by the number of eligible patients available during the study period. A total of 345 patients with suspected CD or ITB were initially identified during model development. After eligibility assessment and exclusion of patients with incomplete clinical data or other reasons for ineligibility, 132 patients were included in model development, comprising 82 with CD and 50 with ITB. For same-center temporal validation, 172 patients with CD and 169 with ITB were initially identified; after screening, 56 patients with CD and 61 with ITB were included. At the independent external validation center, 163 patients with CD and 59 with ITB were initially identified, and 40 patients with CD and 25 with ITB were ultimately included. The screening process for all three cohorts is shown in Supplementary Fig. 1, Supplementary Data Content 1, which details participant screening and exclusions.

Reference Diagnosis, Inclusion Criteria, and Exclusion Criteria

Diagnostic Criteria

No single diagnostic gold standard is available for CD. In accordance with the ECCO-ESGAR guideline and its updated recommendations for the diagnosis of inflammatory bowel disease [28–29], the reference diagnosis integrated the clinical course, laboratory findings, ileocolonoscopic and biopsy histopathologic findings, imaging findings, treatment response, and follow-up data. For typical cases that clearly met guideline-based criteria, the final diagnosis and treatment plan were jointly determined by the treating gastroenterologist and a senior gastroenterologist. Atypical or diagnostically difficult cases were reviewed at a department conference or multidisciplinary consultation.

ITB was classified as microbiologically or histopathologically confirmed when at least one of the following was present: (1) positive acid-fast bacillus staining of lesion tissue or a relevant specimen; (2)

positive culture of Mycobacterium tuberculosis from tissue; (3) a positive Xpert MTB/RIF, Xpert MTB/RIF Ultra, or other nucleic acid test for M. tuberculosis in intestinal or other involved tissue; or (4) histopathologic evidence of typical caseating necrotizing granulomas. In the absence of these confirmatory findings, a clinical diagnosis of ITB required improvement in clinical symptoms after standard antituberculosis therapy, improvement of endoscopic lesions on repeat examination, and no evidence of relapse or progression during 9–12 months of follow-up after treatment [2–3, 30].

CD was classified as pathologically supported when typical histopathologic findings, noncaseating granulomas, or surgical pathology supported the diagnosis. Otherwise, a clinically adjudicated diagnosis of CD was based on the composite framework above, including exclusion of ITB, improvement after CD-directed treatment, and no new evidence supporting ITB or another disease during follow-up. Patients in both disease groups received disease-specific treatment and were generally reassessed with laboratory, endoscopic, or other objective examinations after 3–6 months to evaluate early treatment response [2–3, 23, 29].

Inclusion Criteria

Patients were required to meet all of the following criteria: (1) a final diagnosis of CD or ITB according to the reference-diagnosis criteria above; (2) availability of the pretreatment clinical data and model predictors required for the study, including age at symptom onset, IGRA, and endoscopic features; (3) completion of colonoscopy before the initiation of antituberculosis therapy, corticosteroids, immunosuppressants, biologic agents, or other treatments that could alter examination findings.

Exclusion Criteria

Patients meeting any of the following criteria were excluded: (1) inability to establish a final diagnosis of either CD or ITB, or coexistence of CD and ITB; (2) receipt of standard antituberculosis therapy, corticosteroids, immunosuppressants, or biologic agents for 1 consecutive week or longer before examination; (3) absence of the pretreatment baseline data required for the study; (4) a history of intestinal resection or another surgical procedure that could alter the endoscopic phenotype.

Candidate Predictors

Candidate variables included demographic characteristics, routine laboratory tests and tuberculosis-related tests, endoscopic features, and histopathologic findings.

Demographic variables included sex and age at symptom onset. Age at symptom onset was defined as the patient's age when symptoms attributable to the disease first appeared, determined retrospectively from the medical history documented at the time of the initial diagnosis.

Laboratory data were collected according to their availability during the initial diagnostic evaluation and mainly included serum biochemical measurements, tuberculosis-related tests, and other tests used in the differential diagnosis. Tuberculosis-related candidate predictors included the interferon-gamma release assay (IGRA), acid-fast bacillus smear microscopy, *M. tuberculosis* culture, Xpert MTB/RIF, *M. tuberculosis* PCR, and other nucleic acid tests for *M. tuberculosis*.

Endoscopic features were coded in a structured manner according to the eight colonoscopic features proposed by Lee et al. for differentiating CD from ITB [7]. Disease extent was recorded across six intestinal segments: the ileocecal region, ascending colon, transverse colon, descending colon, sigmoid colon, and rectum. Lesions involving the ileocecal valve were assigned to the ileocecal region. Features supporting CD included anorectal lesions, longitudinal ulcers, aphthous ulcers, and a cobblestone appearance. Features supporting ITB included involvement of fewer than four intestinal segments, transverse or circumferential ulcers, scars or pseudopolyps, and a patulous ileocecal valve. Each feature was coded as 1 if present and 0 if absent, and the total numbers of features supporting CD and ITB were calculated separately.

To assess the reproducibility of the endoscopic features, 32 cases with complete endoscopic images were selected using a random number generator, including 18 from Shenzhen Third People's Hospital and 14 from Shenzhen Second People's Hospital. A second endoscopist, blinded to the patients' final diagnoses, independently reassessed all eight endoscopic features. Case-level complete agreement, item-level agreement, and Cohen's kappa for each feature were calculated. The second endoscopist's assessments were then entered into the fixed formula to compare individual classifications before and after reassessment.

Model Development and Validation

During model development, demographic characteristics, tuberculosis-related laboratory tests, the eight endoscopic features, the Lee endoscopic score, and histopathologic findings were compared between patients with CD and ITB, with ITB coded as 1 and CD as 0. Candidate predictors were evaluated according to previous evidence, clinical plausibility, univariable associations, data availability, and feasibility of implementation. Multivariable logistic regression models were fitted using different combinations of candidate predictors, and their discrimination was compared. Predictor selection also considered clinical feasibility and model simplicity to avoid unnecessary predictors.

Model equation: $Y = -5.603 + 0.070X_1 - 2.520X_2 + 1.163X_3 + 3.955X_4$; $P(ITB) = 1/[1 + \exp(-Y)]$.

In this equation, X_1 is age at symptom onset; X_2 is the number of endoscopic features supporting CD, with 1 point added for each feature present; X_3 is the number of endoscopic features supporting ITB, with 1 point added for each feature present; and X_4 is the IGRA result, coded as 1 for positive and 0 for negative. The regression coefficient, standard error, odds ratio (OR), and 95% confidence interval (CI) for each predictor are shown in Table 2.

During validation, P(ITB) > 0.359 was classified as ITB and P(ITB) ≤ 0.359 as CD. For each patient, age at symptom onset, IGRA, and the eight endoscopic features were entered into the development formula. All 117 patients in the same-center temporal validation cohort and all 65 patients in the independent external validation cohort were included in baseline comparisons, ROC analysis, classification at the fixed cutoff, calibration assessment, Brier score calculation, and comparisons with individual indicators. Regression coefficients were not re-estimated, variables were not reselected, and the cutoff was not modified during validation.

Statistical Analysis

Continuous variables were summarized as medians (interquartile ranges) and compared using the Mann-Whitney U test. Categorical variables were presented as counts and percentages [n/N (%)] and compared using the Pearson chi-square test, continuity-corrected chi-square test, or Fisher's exact test, as appropriate according to expected cell counts. The outcome was coded as ITB = 1 and CD = 0. Exact 95% CIs for proportions were calculated using the Clopper-Pearson method. In the validation cohorts, discrimination was assessed using receiver operating characteristic (ROC) curves and AUC. AUCs and their 95% CIs, as well as paired differences between the full model and IGRA, the Lee endoscopic score, and individual indicators, were assessed using the DeLong method. At the fixed cutoff, sensitivity, specificity, positive predictive value, negative predictive value, and accuracy were reported with 95% CIs. Paired classification results from the model and the fixed IGRA rule were compared using the exact McNemar test. Calibration assessment included the Brier score, observed-to-expected (O/E) ratio, calibration-in-the-large intercept, calibration slope, and grouped calibration plots. Confidence intervals were estimated from 2,000 bootstrap resamples stratified by outcome. When complete or near-complete separation occurred, the calibration slope was reported as not stably estimable. Except for the prespecified primary model performance measures, all comparisons between indicators were considered exploratory, and no adjustment for multiple comparisons was performed.

The Hosmer-Lemeshow test was used to assess goodness of fit in the development sample. To assess coefficient stability, 1,000 bootstrap resamples were generated, and the original regression coefficients of the final model and bootstrap percentile 95% CIs were reported. This bootstrap procedure did not repeat predictor selection or estimate optimism-corrected model-level performance. The classification cutoff was selected in the development cohort using the Youden index, calculated as sensitivity + specificity − 1.

After model development, the fixed formula, regression coefficients, and original cutoff of 0.359 were applied to the same-center temporal validation and independent external validation cohorts. Model predictions were compared with the final diagnoses to evaluate discrimination, calibration, and classification performance at the fixed cutoff in each validation cohort.

Model development and bootstrap analyses were performed using IBM SPSS Statistics 27.0. Validation cohort analyses were performed using Python 3.12, SciPy 1.17.0, and scikit-learn 1.8.0. All tests were two-sided, and P < 0.05 was considered statistically significant.

Results

Cohort Characteristics

In the development cohort, 63 patients with CD (76.8%) and 28 patients with ITB (56.0%) were male, and the sex distribution differed significantly between the groups (P = 0.012). The median ages at symptom onset were 29.5 years (interquartile range, 24.0–37.0) in patients with CD and 40.0 years (interquartile range, 34.0–51.0) in those with ITB; patients with ITB were older at symptom onset (P < 0.001). IGRA positivity, involvement of fewer than four intestinal segments, transverse or circumferential ulcers, and scars or pseudopolyps were more common in ITB, whereas anorectal lesions, longitudinal ulcers, and a cobblestone appearance were more common in CD (Table 1).

Table 1
Demographic, Laboratory, Endoscopic, and Histopathologic Characteristics of the Development Cohort

Variable	CD (n = 82)	ITB (n = 50)	P value
Patient characteristics			
Age at symptom onset, years	29.5 (24.0–37.0)	40.0 (34.0–51.0)	< 0.001
Male sex, n (%)	63 (76.8)	28 (56.0)	0.012
Laboratory tests			
Xpert MTB/RIF positive, n (%)	0/73 (0.0)	19/40 (47.5)	< 0.001
IGRA positive, n (%)	11 (13.4)	40 (80.0)	< 0.001
M. tuberculosis nucleic acid test positive, n (%)	0/51 (0.0)	11/47 (23.4)	< 0.001
Acid-fast bacillus smear positive, n (%)	0/16 (0.0)	5/26 (19.2)	0.138
M. tuberculosis PCR positive, n (%)	0/1 (0.0)	3/9 (33.3)	1.000
M. tuberculosis culture positive, n (%)	0/29 (0.0)	13 (26.0)	0.002
Endoscopic features			
Longitudinal ulcers, n (%)	17 (20.7)	0 (0.0)	< 0.001
Aphthous ulcers, n (%)	5 (6.1)	0 (0.0)	0.156
Cobblestone appearance, n (%)	16 (19.5)	0 (0.0)	< 0.001
Anorectal lesions, n (%)	45 (54.9)	6 (12.0)	< 0.001
Transverse/circumferential ulcers, n (%)	5 (6.1)	10 (20.0)	0.015
Scars or pseudopolyps, n (%)	14 (17.1)	21 (42.0)	0.002
Patulous ileocecal valve, n (%)	6 (7.3)	8 (16.0)	0.116
Involvement of fewer than four intestinal segments, n (%)	47 (57.3)	45 (90.0)	< 0.001
Histopathology			
Granuloma, n (%)	19/78 (24.4)	27/46 (58.7)	< 0.001

Note: Categorical variables are presented as n (%); when data were missing, they are presented as n/N (%), where N is the number of patients who underwent the corresponding test.

Final Model

Xpert MTB/RIF, *M.tuberculosis* nucleic acid testing, acid-fast bacillus smear microscopy, *M.tuberculosis* PCR, and *M.tuberculosis* culture were also evaluated as candidate indicators during model development. Acid-fast bacillus smear microscopy and *M. tuberculosis* PCR did not differ significantly between the CD and ITB groups (P = 0.138 and P = 1.000). Although Xpert MTB/RIF, *M. tuberculosis* nucleic acid testing, and *M. tuberculosis* culture differed significantly between the CD and ITB groups (all P ≤ 0.002), adding them to the candidate models did not further increase the AUC, and they were therefore not retained in the final model. Among the candidate models, the model comprising age at symptom onset, IGRA, the number of endoscopic features supporting CD, and the number supporting ITB yielded the highest AUC and was selected as the final model (Table 2).

For each 1-year increase in age at symptom onset, the odds of ITB increased by 7.3% (OR, 1.073; 95% CI, 1.005–1.145). A positive IGRA result was associated with an OR of 52.217 (95% CI, 11.315-240.972). Each additional endoscopic feature supporting CD was associated with lower odds of ITB (OR, 0.080; 95% CI, 0.014–0.474), whereas each additional feature supporting ITB was associated with higher odds of ITB (OR, 3.198; 95% CI, 1.027–9.960). The AUC of the development model was 0.970 (95% CI, 0.947–0.994), and the Hosmer-Lemeshow test yielded P = 0.667. Results from 1,000 bootstrap resamples showed that the percentile 95% CIs for age at symptom onset, IGRA, and the number of endoscopic features supporting CD did not cross zero, whereas the interval for the number supporting ITB crossed zero (-0.350 to 3.217), indicating comparatively limited stability of this coefficient. Because complete or near-complete separation may have occurred in some resamples, the bootstrap coefficient distributions were markedly skewed. The results were therefore interpreted primarily using the percentile 95% CIs (Supplementary Table 1).

Table 2
Results of the Multivariable Logistic Regression Model

Variable	B	P value	OR	Lower 95% CI	Upper 95% CI	Standard error
Age at symptom onset	0.070	0.036	1.073	1.005	1.145	0.033
IGRA positive	3.955	< 0.001	52.217	11.315	240.972	0.780
CD-supporting endoscopic feature count	-2.520	0.005	0.080	0.014	0.474	0.905
ITB-supporting endoscopic feature count	1.163	0.045	3.198	1.027	9.960	0.580
Intercept	-5.603	0.001	0.004	—	—	1.688

Note: The outcome was coded as ITB = 1 and CD = 0. B denotes the unstandardized regression coefficient; OR, odds ratio; CI, confidence interval.

Model Performance in the Validation Cohorts

In the same-center temporal validation cohort, patients with ITB were older at symptom onset and more frequently had a positive IGRA result, transverse or circumferential ulcers, and involvement of fewer than four intestinal segments. Longitudinal ulcers, aphthous ulcers, a cobblestone appearance, and anorectal lesions were more common in patients with CD. Patients with CD had more endoscopic features supporting CD, whereas patients with ITB had more features supporting ITB. A broadly similar pattern was observed in the independent external validation cohort (Table 3).

Table 3
Main Characteristics of the Validation Cohorts

Variable	Same-center CD (n = 56)	Same-center ITB (n = 61)	P value	External CD (n = 40)	External ITB (n = 25)	P value
Age at symptom onset, years	27.5 (24–34)	37 (29–50)	< 0.001	25.5 (20.75–32.25)	47 (37–55)	< 0.001
Male sex, n/N (%)	45/56 (80.4)	35/61 (57.4)	0.008	31/40 (77.5)	12/25 (48.0)	0.014
IGRA positive, n/N (%)	2/56 (3.6)	59/61 (96.7)	< 0.001	5/40 (12.5)	25/25 (100.0)	< 0.001
Longitudinal ulcers, n/N (%)	14/56 (25.0)	2/61 (3.3)	< 0.001	9/40 (22.5)	0/25 (0.0)	0.029
Aphthous ulcers, n/N (%)	35/56 (62.5)	17/61 (27.9)	< 0.001	40/40 (100.0)	19/25 (76.0)	0.005
Cobblestone appearance, n/N (%)	11/56 (19.6)	0/61 (0.0)	< 0.001	15/40 (37.5)	1/25 (4.0)	0.002
Anorectal lesions, n/N (%)	38/56 (67.9)	5/61 (8.2)	< 0.001	30/40 (75.0)	1/25 (4.0)	< 0.001
Transverse/circumferential ulcers, n/N (%)	1/56 (1.8)	15/61 (24.6)	< 0.001	3/40 (7.5)	9/25 (36.0)	0.011
Scars or pseudopolyps, n/N (%)	15/56 (26.8)	22/61 (36.1)	0.281	17/40 (42.5)	12/25 (48.0)	0.664
Patulous ileocecal valve, n/N (%)	1/56 (1.8)	8/61 (13.1)	0.051	0/40 (0.0)	12/25 (48.0)	< 0.001
Involvement of fewer than four intestinal segments, n/N (%)	19/56 (33.9)	58/61 (95.1)	< 0.001	6/40 (15.0)	22/25 (88.0)	< 0.001
CD-supporting endoscopic feature count	2 (1–3)	0 (0–1)	< 0.001	2 (2–3)	1 (1–1)	< 0.001
ITB-supporting endoscopic feature count	1 (0–1)	2 (1–2)	< 0.001	1 (0–1)	2 (2–2)	< 0.001

In the same-center temporal validation cohort of 61 patients with ITB and 56 with CD, the model correctly classified 58 patients with ITB and all patients with CD. Three patients with ITB were classified as CD, and no patient with CD was classified as ITB. The AUC was 0.9994 (DeLong 95% CI, 0.9980-1.0000),

sensitivity was 95.1% (95% CI, 86.3%-99.0%), specificity was 100.0% (95% CI, 93.6%-100.0%), positive predictive value was 100.0% (95% CI, 93.8%-100.0%), negative predictive value was 94.9% (95% CI, 85.9%-98.9%), and accuracy was 97.4% (95% CI, 92.7%-99.5%).

In the independent external validation cohort, the model correctly classified 22 patients with ITB and all patients with CD; three patients with ITB were classified as CD. The AUC was 0.9995 (DeLong 95% CI, 0.9981-1.0000), sensitivity was 88.0% (95% CI, 68.8%-97.5%), specificity was 100.0% (95% CI, 91.2%-100.0%), positive predictive value was 100.0% (95% CI, 84.6%-100.0%), negative predictive value was 93.0% (95% CI, 80.9%-98.5%), and accuracy was 95.4% (95% CI, 87.1%-99.0%) (Table 4 and Fig. 1).

Table 4
Diagnostic Performance of the Model in the Validation Cohorts

Indicator	**Same-center temporal validation (n = 117)**	**Independent external validation (n = 65)**
Cohort composition	ITB, 61; CD, 56	ITB, 25; CD, 40
AUC (95% CI)	0.9994 (0.9980-1.0000)	0.9995 (0.9981-1.0000)
Sensitivity (95% CI)	95.1% (86.3%-99.0%)	88.0% (68.8%-97.5%)
Specificity (95% CI)	100.0% (93.6%-100.0%)	100.0% (91.2%-100.0%)
Positive predictive value (95% CI)	100.0% (93.8%-100.0%)	100.0% (84.6%-100.0%)
Negative predictive value (95% CI)	94.9% (85.9%-98.9%)	93.0% (80.9%-98.5%)
Accuracy (95% CI)	97.4% (92.7%-99.5%)	95.4% (87.1%-99.0%)

Note: AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value.

Comparison with Individual Indicators

In the same-center temporal validation cohort, the AUCs of the full model, IGRA, the Lee endoscopic score, age at symptom onset, the number of endoscopic features supporting CD, and the number supporting ITB were 0.9994, 0.9657, 0.9486, 0.7285, 0.8431, and 0.8293, respectively. The AUC of the full model was higher than that of IGRA (DeLong P = 0.044) and the Lee endoscopic score (P = 0.002). In the independent external validation cohort, the corresponding AUCs were 0.9995, 0.9375, 0.9875, 0.8870, 0.9370, and 0.9370, respectively. The AUC of the full model was higher than that of IGRA (P = 0.018) but did not differ significantly from that of the Lee endoscopic score (P = 0.218) (Supplementary Table 3 and Fig. 1).

When IGRA positivity was used to classify patients as having ITB, sensitivity, specificity, and accuracy were 96.7%, 96.4%, and 96.6%, respectively, in the same-center temporal validation cohort and 100.0%, 87.5%, and 92.3%, respectively, in the independent external validation cohort. By comparison, the full model had a specificity of 100.0% in both cohorts and accuracies of 97.4% and 95.4%, but sensitivities of 95.1% and 88.0%, respectively. Compared with IGRA alone, the full model increased specificity but reduced sensitivity. Exact McNemar tests comparing paired classifications at the fixed cutoff showed no significant difference in either cohort (P = 1.000 and P = 0.727). Notably, all seven IGRA-positive patients with CD across the two cohorts were correctly classified by the model.

Calibration Performance

In the same-center temporal validation cohort, the Brier score was 0.0476 (95% CI, 0.0297–0.0691), the calibration-in-the-large intercept was 1.930 (95% CI, 1.497–2.457), and the calibration slope was 4.009 (95% CI, 2.721–5.336). In the independent external validation cohort, the Brier score was 0.0424 (95% CI, 0.0139–0.0779), the calibration-in-the-large intercept was 2.473 (95% CI, 1.651–3.485), and the slope could not be stably estimated because of near-complete separation. The positive calibration-in-the-large intercepts in both cohorts indicated that the original model generally underestimated the probability of ITB. A high AUC does not necessarily indicate that individual predicted probabilities are adequately calibrated (Supplementary Fig. 2 and Supplementary Table 2).

Reproducibility of Endoscopic Assessment

Among the 32 cases included in the blinded reassessment, complete agreement at the case level was 78.1% (25/32; 95% CI, 60.0%-90.7%), item-level agreement was 97.3% (249/256; 95% CI, 94.4%-98.9%), and Cohen's kappa values for individual features ranged from 0.671 to 1.000. Entering the second endoscopist's assessments into the fixed formula changed two classifications from CD to ITB. Classification accuracy in the reassessment sample was 90.6% (29/32) using the original assessments and 96.9% (31/32) using the reassessed features.

Discussion

The treatment strategies and disease outcomes of CD and ITB differ substantially, making early and reliable differentiation directly relevant to treatment selection and patient safety. In difficult cases without sufficient microbiologic or characteristic histopathologic evidence, a diagnostic trial of antituberculosis therapy is still sometimes included in the diagnostic process. However, this approach may cause drug-related adverse events, delay treatment for CD, and yield nonspecific short-term symptomatic improvement; its value as the sole basis for diagnosis is therefore limited [2–3, 30]. The difficulty in differentiating CD from ITB does not arise from a lack of potential indicators, but from overlap and conflict among different sources of evidence. The value of a clinical prediction model therefore depends not only on overall discrimination but also on its ability to integrate these findings across different case spectra. Using information routinely available during the initial diagnostic evaluation, we developed a parsimonious model comprising age at symptom onset, IGRA, and two

structured endoscopic feature counts. The model maintained high discrimination in both the same-center temporal and independent external validation cohorts, while providing an interpretable framework for integrating demographic, immunological, and endoscopic information during clinical evaluation.

Interpretation of the Predictors

Age at symptom onset, IGRA, and endoscopic morphology represent the disease spectrum, evidence of tuberculosis infection, and the intestinal mucosal phenotype, respectively. Previous Korean and Chinese multicenter studies both found that patients with ITB were older than those with CD and incorporated age into their prediction models [11, 14]. An Asian multicenter external validation study also suggested that age at symptom onset may provide important information for improving the discrimination of basic models [17]. Although age at symptom onset is not disease-specific, when combined with other predictors it may alter the prior probability for patients with otherwise similar IGRA and endoscopic findings.

In addition to IGRA, Xpert MTB/RIF, M. tuberculosis nucleic acid testing, acid-fast bacillus smear microscopy, M. tuberculosis PCR, and M. tuberculosis culture were examined in the development cohort, although they were not uniformly available. A positive microbiologic result provides strong support for ITB, but meta-analyses have reported a pooled sensitivity of only 23% and a specificity of 100% for Xpert MTB/RIF on intestinal tissue, and a pooled sensitivity of 47% and specificity of 95% for MTB-PCR. Their principal limitation is therefore low sensitivity, and a negative result does not exclude ITB [2, 8–9]. IGRA was available for all 132 patients included in the final development analysis and is relatively standardized and readily available during the initial diagnostic evaluation. A meta-analysis of 11 studies involving 1,081 Asian participants reported a pooled sensitivity of 81%, specificity of 85%, and AUC of approximately 0.92 for IGRA in differentiating ITB from CD [31]. In a prospectively and consecutively enrolled Chinese study, T-SPOT.TB had a sensitivity of 84.2%, specificity of 75.4%, and negative predictive value of 94.2% [32]. In the present study, the AUCs of IGRA alone were 0.9657 and 0.9375 in the same-center temporal and independent external validation cohorts, respectively, further supporting its discriminatory value. However, a study conducted in a high-tuberculosis-burden setting reported a sensitivity of 40.7%, specificity of 75.5%, and AUC of 0.66 for IGRA [33], indicating that performance may vary across populations and clinical settings. IGRA reflects the immune response to M. tuberculosis, but it does not identify the site of infection or reliably distinguish latent infection from active ITB. It should therefore be interpreted together with other evidence and should not be used alone to confirm or exclude ITB [31–33].

Endoscopy plays an important role in differentiating CD from ITB. Longitudinal ulcers, aphthous ulcers, a cobblestone appearance, and anorectal lesions favor CD, whereas transverse or circumferential ulcers, limited disease extent, scars or pseudopolyps, and a patulous ileocecal valve favor ITB. These patterns are consistent with the general direction of the Lee endoscopic score and subsequent studies [7, 10–12, 15]. In this study, the numbers of endoscopic features supporting CD and ITB were calculated separately, and logistic regression was used to estimate the relative weight of each count. This approach retained

the convenience of structured recording, avoided an indeterminate classification when the two feature counts were equal, and produced a continuous probability of ITB. The bootstrap interval for the coefficient of the ITB-supporting feature count crossed zero, indicating that this coefficient requires further evaluation in a larger sample.

Comparison with Other Diagnostic Models

The discriminatory performance of individual clinical, laboratory, or endoscopic indicators is generally limited and may vary across case spectra. Combining several important parameters in a multivariable model can preserve the direction of each source of evidence while producing an individualized continuous probability [10–17]. Makharia et al. incorporated clinical, endoscopic, and histopathologic features into a weighted score. Jung et al. subsequently developed a seven-marker model that included age, symptoms, ulcer morphology, and radiologic evidence of pulmonary tuberculosis. Both models showed good discriminatory performance in their original studies, but later validation suggested that performance could decline with changes in the case spectrum and study center [10–11, 17]. Bae et al. added ASCA, IGRA, proximal small-bowel involvement, and radiologic evidence of pulmonary tuberculosis to the Lee endoscopic score, increasing diagnostic accuracy in the development cohort from 81.2% to 96.3% and demonstrating that information from different diagnostic modalities may complement one another [12]. In a prospective six-center study, He et al. developed a nomogram incorporating age, IGRA, endoscopic findings, and computed tomography enterography features and validated it in an independent cohort. However, the model depends to some extent on complete imaging and specialist interpretation [14]. Lu et al. used a classification and regression tree approach to convert IGRA, disease extent, and ulcer morphology into an intuitive decision pathway. Zeng et al. developed a nomogram combining T-SPOT.TB, endoscopic, radiologic, and histopathologic features, but relied mainly on internal validation and lacked independent external validation [15–16]. Notably, Limsrivilai et al. externally validated several previous models in the same cohort across five Asian centers and reported AUROCs of 0.862 for the Jung model and 0.853 for a clinical-endoscopic model, whereas the AUROC of the Lee endoscopic score was only 0.713 [17]. Although some studies have performed temporal or independent-cohort validation, cross-center studies indicate that both AUC and classification performance at the original cutoff may still decline [12, 14–15, 17].

Together, these studies show that multivariable integration may address some limitations of individual indicators. However, as the number of predictors increases, a model becomes more dependent on complete testing, equipment, and specialist interpretation, which may limit its use in settings with different resources. The present study used four routinely available predictors to construct a parsimonious and clinically interpretable model and assessed its performance across different time periods and centers through same-center temporal validation and independent external validation (Supplementary Table 4).

Misclassification by the Model

The clinical relevance of a model depends not only on its AUC but also on which patients are misclassified. These false-negative classifications are clinically important because inappropriate immunosuppressive treatment in unrecognized ITB may lead to adverse consequences. Across the validation cohorts, six patients with ITB were classified as having CD. Some IGRA-positive patients with ITB fell below the cutoff because of a younger age at symptom onset, a greater number of CD-supporting endoscopic features, or fewer ITB-supporting features. This indicates that the model can integrate conflicting information but may assign a low probability to patients with atypical ITB. A predicted probability below the cutoff should therefore not be used as a basis for discontinuing microbiologic testing for tuberculosis, chest or abdominal imaging, or specialist assessment, or for initiating immunosuppressive therapy.

Reproducibility of Endoscopic Interpretation

Structured endoscopic features are important model inputs and are also the component most likely to be influenced by observer experience. Among the 32 cases included in the blinded reassessment showed high overall item-level agreement, but small differences in feature interpretation moved two patients across the fixed classification threshold. This finding highlights the need for standardized definitions, structured endoscopic assessment, and reader training before broader implementation.

Strengths and Limitations

The main strengths of this study include the use of only four commonly available predictors, with an explicit formula and cutoff. In addition to overall performance, we reported cases misclassified at the fixed cutoff and conducted a second blinded image review to provide an initial assessment of the reproducibility of the endoscopic inputs. We also performed same-center temporal validation and independent external validation, providing preliminary evidence of temporal stability and transportability across centers.

This study also has several limitations. First, complete-case analysis was used during model development, which may have introduced selection bias and shifted the case spectrum. Second, candidate predictors were evaluated in the development cohort, and predictor selection was not repeated within the bootstrap resamples; overfitting and performance overestimation remain possible [34]. IGRA and endoscopic findings may have contributed to the final diagnostic assessment while also serving as model predictors; incorporation bias and review bias therefore cannot be completely excluded. Third, the validation cohorts consisted of patients with established final diagnoses and may not represent the full case spectrum of consecutively presenting patients requiring initial differentiation. The independent external validation cohort was relatively small and derived from a single external center, and both centers were located in the same region. Future studies should prespecify the sample size for external validation and assess the model in prospectively and consecutively enrolled multicenter cohorts across different geographic regions and levels of care.

Conclusions

We developed a parsimonious model comprising age at symptom onset, IGRA, the number of endoscopic features supporting CD, and the number supporting ITB. The model showed promising discrimination and classification performance in the validation cohorts at the fixed cutoff in the same-center temporal and independent external validation cohorts and correctly classified some cases with discordant tuberculosis-related test results and endoscopic findings. The model may serve as an adjunctive stratification tool within a comprehensive diagnostic pathway, supporting decisions regarding further microbiologic sampling, imaging assessment, specialist consultation, or multidisciplinary discussion. Its probability calibration, clinical net benefit, and generalizability across regions require further evaluation in prospectively and consecutively enrolled multicenter cohorts.

Declarations

Acknowledgments: Artificial intelligence-assisted tools were used for language editing and translation assistance. The authors reviewed and approved all content and take full responsibility for the accuracy and integrity of the manuscript.

Human Ethics: This study was approved by the Scientific Research Ethics Committee of Shenzhen Third People's Hospital (approval number: 2026-140-01). The study was conducted in accordance with applicable institutional and ethical standards.

Consent to Participate: The requirement for informed consent to participate was waived by the Scientific Research Ethics Committee of Shenzhen Third People's Hospital because this was a retrospective study using de-identified clinical data.

Author Contributions: Xiaoyin Zhang conceived and supervised the study. Ruibin Xu, Chuyao Zhang, and Weitao Wu participated in data collection. Ruibin Xu and Chuyao Zhang performed statistical analyses and model development. Weitao Wu contributed to external validation, which was supervised by Dayong Sun. Ruibin Xu drafted the manuscript and prepared the main text and figures. All authors reviewed and approved the final manuscript.

Funding: None.

Conflicts of Interest: The authors declare no conflicts of interest related to this study.

Data Availability: The data supporting the conclusions of this study are not publicly available because of patient privacy and institutional ethics restrictions. Access may be requested from the corresponding author upon reasonable request and is subject to approval by the participating institutions and ethics committees.

Financial Disclosures: The authors have no financial disclosures related to this work.

References

1. Kedia S, Das P, Madhusudhan KS, et al. Differentiating Crohn's disease from intestinal tuberculosis. World J Gastroenterol. 2019;25(4):418–32. 10.3748/wjg.v25.i4.418.
2. Choudhury A, Dhillon J, Sekar A, et al. Differentiating gastrointestinal tuberculosis and Crohn's disease: a comprehensive review. BMC Gastroenterol. 2023;23(1):246. 10.1186/s12876-023-02887-0.
3. Pratap Mouli V, Kedia S, Sharma V, et al. Navigating the differential diagnosis of intestinal tuberculosis and Crohn's disease. Gastroenterology. 2025;169(6):1139–46. 10.1053/j.gastro.2025.08.007.
4. Ng SC, Tang W, Ching JY, et al. Incidence and phenotype of inflammatory bowel disease based on results from the Asia-Pacific Crohn's and Colitis Epidemiology Study. Gastroenterology. 2013;145(1):158–e1652. 10.1053/j.gastro.2013.04.007.
5. World Health Organization. Global tuberculosis report 2025. Geneva: World Health Organization; 2025.
6. Seo H, Lee S, So H, et al. Temporal trends in the misdiagnosis rates between Crohn's disease and intestinal tuberculosis. World J Gastroenterol. 2017;23(34):6306–14. 10.3748/wjg.v23.i34.6306.
7. Lee YJ, Yang SK, Byeon JS, et al. Analysis of colonoscopic findings in the differential diagnosis between intestinal tuberculosis and Crohn's disease. Endoscopy. 2006;38(6):592–7. 10.1055/s-2006-924996.
8. Sharma V, Soni H, Kumar MP, et al. Diagnostic accuracy of the Xpert MTB/RIF assay for abdominal tuberculosis: a systematic review and meta-analysis. Expert Rev Anti Infect Ther. 2021;19(2):253–65. 10.1080/14787210.2020.1816169.
9. Jin T, Fei B, Zhang Y, He X. The diagnostic value of polymerase chain reaction for Mycobacterium tuberculosis to distinguish intestinal tuberculosis from Crohn's disease: a meta-analysis. Saudi J Gastroenterol. 2017;23(1):3–10. 10.4103/1319-3767.199135.
10. Makharia GK, Srivastava S, Das P, et al. Clinical, endoscopic, and histological differentiations between Crohn's disease and intestinal tuberculosis. Am J Gastroenterol. 2010;105(3):642–51. 10.1038/ajg.2009.585.
11. Jung Y, Hwangbo Y, Yoon SM, et al. Predictive factors for differentiating between Crohn's disease and intestinal tuberculosis in Koreans. Am J Gastroenterol. 2016;111(8):1156–64. 10.1038/ajg.2016.212.
12. Bae JH, Park SH, Ye BD, et al. Development and validation of a novel prediction model for differential diagnosis between Crohn's disease and intestinal tuberculosis. Inflamm Bowel Dis. 2017;23(9):1614–23. 10.1097/MIB.0000000000001162.
13. Limsrivilai J, Shreiner AB, Pongpaibul A, et al. Meta-analytic Bayesian model for differentiating intestinal tuberculosis from Crohn's disease. Am J Gastroenterol. 2017;112(3):415–27. 10.1038/ajg.2016.529.
14. He Y, Zhu ZH, Chen YJ, et al. Development and validation of a novel diagnostic nomogram to differentiate between intestinal tuberculosis and Crohn's disease: a 6-year prospective multicenter

study. Am J Gastroenterol. 2019;114(3):490–9. 10.14309/ajg.0000000000000064.

15. Lu Y, Chen YH, Peng X, et al. Development and validation of a new algorithm model for differential diagnosis between Crohn's disease and intestinal tuberculosis: a combination of laboratory, imaging and endoscopic characteristics. BMC Gastroenterol. 2021;21:291. 10.1186/s12876-021-01838-x.
16. Zeng SX, Lin Y, Guo JX, et al. Differential diagnosis of Crohn's disease and intestinal tuberculosis: development and assessment of a nomogram prediction model. BMC Gastroenterol. 2022;22:461. 10.1186/s12876-022-02519-z.
17. Limsrivilai J, Lee CK, Prueksapanich P, et al. Validation of models using basic parameters to differentiate intestinal tuberculosis from Crohn's disease: a multicenter study from Asia. PLoS ONE. 2020;15(11):e0242879. 10.1371/journal.pone.0242879.
18. Weng F, Meng Y, Lu F, et al. Differentiation of intestinal tuberculosis and Crohn's disease through an explainable machine learning method. Sci Rep. 2022;12:1714. 10.1038/s41598-022-05571-7.
19. Wu X, Huang H, Hou H, et al. Diagnostic performance of a 5-marker predictive model for differential diagnosis between intestinal tuberculosis and Crohn's disease. Inflamm Bowel Dis. 2018;24(11):2452–60. 10.1093/ibd/izy154.
20. Zhao XS, Wang ZT, Wu ZY, et al. Differentiation of Crohn's disease from intestinal tuberculosis by clinical and CT enterographic models. Inflamm Bowel Dis. 2014;20(5):916–25. 10.1097/MIB.0000000000000025.
21. Mao R, Liao WD, He Y, et al. Computed tomographic enterography adds value to colonoscopy in differentiating Crohn's disease from intestinal tuberculosis: a potential diagnostic algorithm. Endoscopy. 2015;47(4):322–9. 10.1055/s-0034-1391230.
22. Huang X, Liao WD, Yu C, et al. Differences in clinical features of Crohn's disease and intestinal tuberculosis. World J Gastroenterol. 2015;21(12):3650–6. 10.3748/wjg.v21.i12.3650.
23. Narang H, Kedia S, Ahuja V. New diagnostic strategies to distinguish Crohn's disease and gastrointestinal tuberculosis. Curr Opin Infect Dis. 2024;37(5):392–401. 10.1097/QCO.0000000000001054.
24. Zhang F, Xu C, Ning L, et al. Exploration of serum proteomic profiling and diagnostic model that differentiate Crohn's disease and intestinal tuberculosis. PLoS ONE. 2016;11(12):e0167109. 10.1371/journal.pone.0167109.
25. Zhang S, Luo J, Wu Z, et al. Antibodies against glycoprotein 2 display diagnostic advantages over ASCA in distinguishing Crohn's disease from intestinal tuberculosis and intestinal Behçet's disease. Clin Transl Gastroenterol. 2018;9(2):e133. 10.1038/ctg.2018.1.
26. Tiwari V, Kedia S, Garg SK, et al. CD4 + CD25+FOXP3 + T cell frequency in the peripheral blood is a biomarker that distinguishes intestinal tuberculosis from Crohn's disease. PLoS ONE. 2018;13(2):e0193433. 10.1371/journal.pone.0193433.
27. Roy S, Ghosh S, Banerjee M, et al. A combination of circulating microRNA-375-3p and chemokines CCL11, CXCL12, and G-CSF differentiate Crohn's disease and intestinal tuberculosis. Sci Rep. 2021;11(1):23303. 10.1038/s41598-021-02383-z.

28. 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 Crohns Colitis. 2019;13(2):144–64. 10.1093/ecco-jcc/jjy113.
29. Kucharzik T, Taylor S, Allocca M, et al. ECCO-ESGAR-ESP-IBUS Guideline on diagnostics and monitoring of patients with inflammatory bowel disease: Part 1. J Crohns Colitis. 2025;19(7):jjaf106. 10.1093/ecco-jcc/jjaf106.
30. Pratap Mouli V, Munot K, Ananthakrishnan A, et al. Endoscopic and clinical responses to anti-tubercular therapy can differentiate intestinal tuberculosis from Crohn's disease. Aliment Pharmacol Ther. 2017;45(1):27–36. 10.1111/apt.13840.
31. Ng SC, Hirai HW, Tsoi KK, et al. Systematic review with meta-analysis: accuracy of interferon-gamma release assay and anti-Saccharomyces cerevisiae antibody in differentiating intestinal tuberculosis from Crohn's disease in Asians. J Gastroenterol Hepatol. 2014;29(9):1664–70. 10.1111/jgh.12645.
32. Li Y, Zhang LF, Liu XQ, et al. The role of in vitro interferon-gamma release assay in differentiating intestinal tuberculosis from Crohn's disease in China. J Crohns Colitis. 2012;6(3):317–23. 10.1016/j.crohns.2011.09.002.
33. Sachdeva K, Kumar P, Kante B, et al. Interferon-gamma release assay has poor diagnostic accuracy in differentiating intestinal tuberculosis from Crohn's disease in tuberculosis endemic areas. Intest Res. 2023;21(2):226–34. 10.5217/ir.2022.00010.
34. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004;66(3):411–21. 10.1097/01.psy.0000127692.23278.a9.

Supplementary Files

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

- SupplementaryMaterial.docx