

Application of Machine Learning to Distinguish IBS from Crohn's Disease in Underdeveloped Regions

Varun Bhargav Kurra¹, Microbiome Cornell PHD Jennifer Otiono¹

Received May 11, 2026

Accepted September 16, 2026

Electronic access October 31, 2026

Background and Aims: Crohn's disease, a chronic inflammatory bowel disease (IBD), and Irritable Bowel Syndrome (IBS) share overlapping symptoms, and approximately 10-30% of patients later diagnosed with Crohn's disease are initially misdiagnosed with IBS. The burden is greatest in low- and lower-middle-income countries (LLMICs), where diagnostic delay is roughly twice that of high-income settings. This study evaluates whether Machine Learning (ML) can separate the two conditions using routinely recorded clinical information.

Methods: A cohort of 4,663 patients (1,772 Crohn's disease, 2,891 IBS) was extracted from MIMIC-IV (v3.1). Patients coded for both conditions, ulcerative colitis, or intestinal infection were excluded. Features were restricted to each patient's index admission. Because discharge summaries can state the diagnosis directly, a leakage-controlled text variant retaining only pre-diagnosis sections was constructed alongside the complete note. Clinical notes were vectorized using ClinicalBERT and TF-IDF, and models were benchmarked against a majority-class baseline and simpler classifiers. The test set was locked before tuning.

Results: The leakage-controlled model achieved an accuracy of 76.67% (95% CI 72.67-80.67), an F1-score of 75.58%, and an AUC of 83.87% (95% CI 79.94-87.61), against a baseline accuracy of 58.44%. Unredacted notes scored 0.087 to 0.107 AUC higher, quantifying recorded diagnostic text. Decision-curve analysis showed positive net benefit, and no misclassification occurred at high confidence.

Conclusions: Diagnostic signal separating Crohn's disease from IBS is recoverable from pre-diagnosis clinical text alone, using a model requiring no laboratory panel, specialist review, or specialized hardware. These retrospective single-dataset findings establish feasibility and support external validation in LLMIC settings.

Keywords: Crohn's Disease, Irritable Bowel Syndrome (IBS), Artificial Intelligence (AI), Machine Learning (ML), Clinical Notes, Laboratory Data, Diagnostic Triage

Introduction

Crohn's disease, one of the two main forms of IBD, and Irritable Bowel Syndrome share common symptoms, and both affect one's digestive tract, making maintaining diagnostic accuracy challenging.¹ Reflecting this difficulty, average diagnostic delays, meaning the interval before a correct diagnosis is reached after misdiagnosis, range from 2 months to 8 years in IBD, with Crohn's disease averaging longer delays than Ulcerative Colitis, the second IBD disease.² The diagnostic delay of Crohn's disease leads to the necessity of stronger medication, hospital admissions, and expensive surgeries, imposing heavy cost burdens on those who cannot afford them.³

This research investigated whether AI models leveraging clinical notes and laboratory data can accurately distinguish between IBS and Crohn's disease to improve diagnostic accuracy and reduce misdiagnosis. The hypothesis is that integrating clinical notes and laboratory data will separate the two

conditions more accurately than either data source alone and more accurately than a majority-class baseline.

A second objective emerged as central to this work. Discharge summaries are written after a diagnosis has been reached and routinely state that diagnosis explicitly, which means a model trained on such text can report high accuracy while recovering the recorded answer rather than the underlying clinical pattern. This study therefore treats the separation of genuine diagnostic signal from recorded diagnostic text as a primary objective and reports performance with and without leakage control. Quantifying this effect is rarely done in clinical natural language processing, and doing so here strengthens the evidence that the reported performance reflects clinical information.

The scope of this study is a retrospective analysis of a single database drawn from one United States academic medical center. It does not include prospective patients and does not include data from any LLMIC. The LLMIC burden described below motivates the research question and defines the setting

¹ Parkway West Highschool, Missouri, USA

the work is directed toward. What this study establishes is whether the signal needed for such a tool exists in routinely collected data, which is the necessary first step toward building one. The findings of this study could inform healthcare systems, contribute to clinical guidelines, and serve as a foundation for the future application of AI-driven predictive diagnostic tools in the field of gastroenterology.

Literature Review

Crohn's Disease and IBS Introduction

Crohn's disease is a chronic autoimmune condition that inflames and irritates one's digestive tract. Crohn's disease and Ulcerative Colitis are the 2 primary forms of inflammatory bowel disease (IBD). Symptoms of Crohn's disease may develop gradually or suddenly, and may be mild or severe. A "flare" is a term used in Crohn's disease to describe when symptoms occur. The etiology of Crohn's disease is unknown, but it is widely believed that genetic predisposition and immune system irregularities contribute to the development of the disease. Crohn's disease affects more than 750,000 people in the United States and approximately 6 to 8 million people globally.⁴

Irritable bowel syndrome (IBS) is a common condition that affects the gastrointestinal tract, the stomach, and the intestines. Symptoms of IBS tend to be present for an extended period, making IBS an ongoing condition that needs long-term care. Similar to Crohn's disease, the causes of IBS are unknown, but factors that reportedly contribute to the syndrome include muscle contractions in the intestine, issues with the nervous system, severe infections, early-life stress, and changes in gut microbes. The "flares" of IBS may be triggered by food and anxiety, although these triggers do not initially cause the symptoms to occur, but rather enhance their effect.⁵ A 2015 clinical review conducted by Chey et al.⁶ reported a global IBS prevalence of approximately 12%, with estimates ranging from 7% to 21% depending on the diagnostic criteria applied. In the United States, IBS is estimated to affect 10% to 15% of adults, of whom only 5% to 7% are initially diagnosed with the syndrome, corresponding to roughly 24 to 45 million affected adults nationally. IBS is also known to affect more women, with twice as many women obtaining the condition as men annually.⁷

IBS and Crohn's disease are both known for their "flare-remission" cycle, with "flare-ups" signifying the period of increased symptom severity. "Flare-ups" are when symptoms of the disease become more noticeable after a period of remission, with its causation relating to a dysfunctional immune response where one's immune system attempts to attack germs that enter the body, but uncontrollably harms bodily tissues instead. Inflammation happens when an autoimmune response

occurs, when the immune system attacks one's body cells inappropriately while attempting to remove germs from the body. Inflammation is a central signal that the attack is ongoing and the flare-up is active. Once the bodily threat is gone, the immune system flows into a state of repair, and inflammation may begin to lessen, but in certain cases, inflammation does not completely heal. The flare-ups in both conditions are typically short-term but vary in duration, lasting a few hours to weeks in IBS and from a few days to months in Crohn's disease.^{8,9} During periods of remission, symptoms are reduced or absent, but the underlying disease is still present.

Although both IBS and Crohn's disease follow a flare-remission pattern, they vary significantly in their nature and underlying fundamentals. Crohn's disease is an organic IBD characterized by visible inflammation, ulceration, and structural tissue damage commonly found in the ileocolon. Unlike Crohn's disease, IBS is a functional disorder, classified as a syndrome, meaning that patients experience real symptoms but no observable damage occurs. Patients with IBS do not experience inflammation or structural bowel damage, which patients with Crohn's disease do. This distinction is the biological basis on which the two conditions are separable in principle, and it is the distinction the models developed in this study attempt to recover from routinely collected data.

Overlapping Symptoms and Delays

IBS and IBD, including Crohn's disease and Ulcerative colitis, have an overlap of apparent symptoms, making patients susceptible to frequent misdiagnosis. The overlapping symptoms occur as both conditions affect the gastrointestinal tract, which has a narrow repertoire of symptoms and a limited range of ways it can manifest disease. As a result, multiple gastrointestinal conditions, such as IBS and Crohn's disease, may appear with the same clinical features. Patients with either disorder may experience changes in bowel habits, typically involving diarrhea along with pain, bloating, and rectal mucus discharge.

A systematic review and meta-analysis by Halpin and Ford analyzed studies of patients with IBD who had answered symptom questionnaires for IBS. Their findings revealed that approximately 35% to 40% of patients with IBD reported IBS-related symptoms, and patients with IBD were roughly 5 times more likely to have symptoms compatible with IBS than healthy controls.¹⁰ Their analysis also concluded that the prevalence of IBS-type symptoms in IBD was higher in Crohn's disease than in Ulcerative colitis, with an odds ratio of 1.62, indicating higher odds of reporting IBS-compatible symptoms among patients with Crohn's disease.

One might assume that the underlying symptoms resolve during remission, allowing for an accurate and clear diagnosis. However, this is not the case, as Dr. Keeley M. Fairbrass of the

Leeds Gastroenterology Institute explains, “IBS-type symptoms co-exist in IBD patients even during periods of deep remission.”¹¹ A 2020 systematic review and meta-analysis conducted by Fairbrass et al.¹¹ of 27 studies (15 reporting Crohn’s disease separately) published between January 1, 2012, and May 11, 2020, which included 3,169 patients with IBD in remission, concluded that IBS-type symptoms among patients with IBD who were in remission occurred in 32.5% of patients (95% CI 27.4-37.9), ranging from 11.2% to 63.6% across individual studies. The analysis also found that among the 15 studies of patients with Crohn’s disease in remission, the prevalence of IBS-type symptoms was 36.6% (95% CI 29.5-44.0), compared with 28.7% (95% CI 22.9-34.8) in Ulcerative colitis, an odds ratio of 1.58 (95% CI 1.27-1.98). In conclusion, IBS-type symptoms occur in approximately 35% to 40% of patients with IBD and in 32.5% of patients with IBD in remission, and they are consistently more prevalent in Crohn’s disease than in Ulcerative colitis, providing a solid basis for frequent misdiagnosis.

With IBS and Crohn’s disease patients obtaining overlapping symptoms, diagnostic delays prevailed. A 2023 meta-analysis conducted by Jayasooriya et al.¹² of 101 EMBASE [Excerpta Medica dataBASE] and Medline studies representing over 112,194 patients with IBD (CD = 59,359; UC = 52,835) concluded that the median time to Crohn’s disease diagnosis was 8 months (Interquartile Range [IQR]: 5 months to 15.2 months). In high-income countries, this was 6.2 months (IQR: 5 months to 12.3 months) compared to 11.7 months (IQR: 8.3 months to 18 months) in low- and middle-income countries. Another study consisting of 46 provider surveys across 15 low- and lower-middle-income countries reported that most patients had symptoms of Crohn’s disease 6 to 24 months before diagnosis, and 26% to 50% of patients lived in rural areas, exacerbating the lack of medical resources and efficiencies in rural, low-income provinces.¹³ In conclusion, patients with Crohn’s disease in low- and lower-middle-income countries experience an average diagnostic delay roughly 2 times longer than in high-income settings.

Misdiagnosis: Crises Between Crohn’s Disease and IBS

The most significant cause of diagnostic delays is widespread misdiagnosis. A 2014 case-controlled study published in the United European Gastroenterology (UEG) journal, consisting of over 20,193 IBD cases, revealed that IBD patients were more than 3 times as likely as controls to have a prior record of IBS. 15% of IBD cases compared to 5% of controls had IBS encoded in their medical records before diagnosis, with 11% of IBD cases (cf. 5% of controls) having the code of IBS for over a year before IBD, and over 6% of cases (cf. 3% of controls) maintaining the code for over 5 years. The study concluded that globally, approximately 10% of IBD patients are

initially misdiagnosed with IBS, and in the case of 3% of patients, the misdiagnosis persisted for 5 or more years.¹⁴ Along with the 10% risk of initial misdiagnosis, a large longitudinal study analyzing 9,341 incident cases found that individuals with IBS had an 8.6-fold increased risk of receiving an IBD diagnosis compared to non-IBS controls (238.1 vs. 27.8 cases per 100,000 person-years, $p < 0.0001$). In a subset of well-defined IBS cases, the risk increased to 15 times that of the non-IBS cases, highlighting the potential of initial misdiagnosis and early undetected IBD among IBS patients.¹⁵ Overall, the global rate of Crohn’s disease to IBS misdiagnosis can be stated as 10%, with 3% of total patients being misdiagnosed for 5 years or more. Additionally, IBS patients have an 8.6 to 15 times higher risk than non-IBS patients of obtaining IBD, highlighting a significant and ongoing misdiagnosis crisis.

Misdiagnosis: Crises In Low- And Lower-Middle-Income Countries

The misdiagnosis crisis is a notable burden in high-income countries. Still, the situation becomes considerably more severe in low- and lower-middle-income countries (LLMICs) due to unaffordable costs and scarcity of diagnostic tools and medical professionals. Although the crisis is exacerbated in LLMICs by a scarcity of equipment and professionals and by unaffordable costs, concrete statistical data on Crohn’s disease misdiagnosis as IBS and diagnostic delay remain unavailable. A scoping review of 4,486 publications conducted by Rajbhandari et al.¹⁶ from 79 LLMICs revealed that only 21 LLMICs have publications on Crohn’s disease, indicating major under-diagnosis or under-reporting and outlining the scarcity of statistical data in LLMICs regarding Crohn’s disease. This absence of surveillance data is itself a component of the crisis, since a misdiagnosis burden that is never measured is also never addressed.

In the absence of directly reported misdiagnosis rates for these regions, an indicative estimate of 33.5% was derived in this study by applying the ratio of recorded Crohn’s disease to IBS presentations in LLMIC settings to the misdiagnosis relationships established in high-income cohorts. This figure is presented as an estimate derived from indirect evidence rather than as a reported measurement, and it should be read as an indication of scale rather than a validated rate. Its magnitude is consistent with the diagnostic capacity data below, which independently indicate a diagnostic environment substantially more constrained than the high-income settings in which the 10% figure was measured.

Factors such as diagnostic affordability and resource availability also make an impact on LLMICs. Firstly, regarding diagnostic cost, a cross-sectional survey of 46 gastroenterology providers across 15 LLMICs in Africa and South Asia, conducted by Rajbhandari et al.¹³, reported that 52.4% of Crohn’s

disease patients in Asia and 60.9% of patients in Africa could not afford diagnostic testing to distinguish between Crohn's disease and other diseases. Secondly, addressing the scarcity of diagnostic resource availability, analyzing Africa's metrics reveals a shortfall in equipment and specialists. It should be noted that the gold-standard method of Crohn's disease diagnosis is endoscopy, including colonoscopy and ileoscopy. A 2024 report on the progress of expanding gastrointestinal endoscopy in LLMIC African nations, conducted by Gaur et al.¹⁷, revealed that Eastern Africa's endoscopy capacity consisted of approximately 0.12 endoscopists, 0.12 gastroscopes, and 0.09 colonoscopes available per 100,000 inhabitants, showcasing a stark deficit of diagnostic equipment. Addressing the aspect of the availability of medical specialties, the provider survey disclosed that in several LLMIC African countries, such as Malawi and Rwanda, there was less than 1 medically trained gastroenterologist in each country, serving more than 10 million inhabitants, and fewer than 11 endoscopy centers.¹³ Similarly, Ruma Rajbhandari, a Harvard professor and lead Crohn's disease researcher, stated that "researchers found 'very few gastroenterologists, less than 20' for Ethiopia, which has a population of more than 115 million."¹⁸ Overall, through a statistical analysis of the quantity of gastroenterologists operating in Africa, a generalizable average gastroenterologist density of 0.005 to 0.02 per 100,000 inhabitants, or 1 specialist per 5 to 20 million inhabitants, can be established. In conclusion, the scarcity of resources and specialists is widely apparent in African LLMIC nations, with 0.12 endoscopists, 0.12 gastroscopes, and 0.09 colonoscopes available per 100,000 inhabitants, and an average gastroenterologist density of 0.005 to 0.02 per 100,000 inhabitants being generalizable metrics between LLMICs. The following scarcity of resources and specialists highlights the leading cause of an inflated misdiagnosis burden between Crohn's disease and IBS in LLMICs.

Predictive AI Introduction

With nations seeking alternative, available, and low-cost diagnostic solutions^{19,20}, Artificial Intelligence's (AI's) established background in predictive healthcare stands as a promising approach to improve early detection in LLMICs. AI has expanded rapidly in healthcare over the past 2 decades, driven by the development of Machine Learning (ML) and Deep Learning (DL), the digitization of healthcare records, and a rise in available medical data. The global AI in healthcare market grew from \$1.1 billion in 2016 to \$22.4 billion in 2023 and is projected to reach \$187.69 billion by 2030.^{21,22} A 2020 bibliometric analysis by Guo et al.²³ found that publication growth in healthcare AI rose to 45.15% between 2014 and 2019, against an average of 17.02% per year since 1995. Predictive AI, which applies ML algorithms to medical data, is

now used in early disease detection, patient outcome prediction, and resource optimization.

The utilization of AI in healthcare allows for increased diagnostic accuracy, cost savings, and accessibility. Economic modeling estimates that the wider adoption of AI in healthcare could lead to savings of 5% to 10% in US healthcare spending, around \$200 to \$360 billion annually.²⁴ Industry analyses have further projected substantial reductions in preventable clinical error through AI-assisted workflows, though these projections are based on modeling rather than on measured clinical outcomes.²⁵ AI can be utilized in regions where medical professionals are scarcely available, especially when it comes to diagnostic specialties. In the cases of many LLMICs, AI-based predictive ML models have been extensively tested and evaluated in the cases of Tuberculosis (TB) and Breast Cancer screenings.

A cost-effectiveness analysis by Nsengiyumva et al.²⁶ evaluated TB triage from chest radiograph interpretation using 5 AI models on 23,954 patients in Dhaka, Bangladesh. Reported performance reached an AUC of up to 90.8% at 90% sensitivity, with qXR specificity of 74.3% and CAD4TB of 72.9%, meeting WHO triage targets. AI-based triage was projected to lower costs by 19% to 37% and to avert 3% to 4% of Disability-Adjusted Life Years (DALYs), signifying substantial economic impact alongside improved diagnostic yield.

Similarly, an observational study by Kakileti et al.²⁷ of Thermalytix for breast cancer detection in 470 women reported an AUC of 0.90, sensitivity of 91.02%, and specificity of 82.39%. A real-world evaluation by Adapa et al.²⁸ across 183 locations in Punjab reported an age-adjusted incidence of 120.94 per 100,000 through Thermalytix screening, a three-fold improvement over the nationally recommended clinical breast examination pathway.

These precedents establish that AI-based triage delivers measurable diagnostic and economic benefit in resource-constrained settings, and both were evaluated prospectively within the target setting. The present study addresses the necessary prior question of whether the signal separating Crohn's disease from IBS is recoverable from routinely collected data at all.

Data & Methodology

Research Design

This study is a retrospective, observational analysis of a single medical database. No prospective patients were enrolled, and no clinical decisions were influenced by the models developed. The unit of analysis is the individual patient rather than the hospital admission.

Database Selection

To build the IBS and Crohn's Disease detection ML model, assessments of medical-grade datasets were conducted to align with rigorous benchmarks. Criteria for the database were established as the industry standards for modern-day datasets. The criteria implemented in the dataset assessments are data (a) quality, (b) structure, (c) usability, and (d) patient privacy.

After the criteria were determined, the web was searched for datasets that not only aligned with the criteria but also had noticeable recognition in the inflammatory bowel disease research sector. The database of MIMIC looked to be the most notable, specifically the dataset MIMIC-IV (v3.1) and its subset, MIMIC-IV-Note (de-identified free-text clinical notes). MIMIC-IV (v3.1) is a large, open-sourced medical database developed by the MIT Laboratory for Computational Physiology (MLCP) in alliance with Beth Israel Deaconess Medical Center (BIDMC). The database contains de-identified data from both intensive care unit (ICU) and medical-surgical care (hospital) patients admitted between 2008 and 2019. The dataset has 364,627 unique patients, 94,458 ICU stays, and 546,028 hospital admissions.

The MIMIC-IV (v3.1) database aligns with all 4 criteria. For quality, procedures and diagnoses are encoded under ICD and CPT standards with consistent validation and longitudinal patient records. For structure, the data are organized into linked tables with labeled metadata using ICD, LOINC, and CPT terminologies. For usability, the database is publicly available, extensively documented by PhysioNet, and compatible with standard Python tooling. For privacy, it follows a regulated de-identification process removing the 18 identifiers mandated by HIPAA, applies date shifting, and requires human research training and a signed data usage agreement. This database met all established criteria and was selected for this research.

It should be stated directly that MIMIC-IV is drawn from a single academic medical center in Boston, United States. Its patient population differs from LLMIC populations in demographics, disease severity, comorbidity burden, healthcare access, and in the completeness of the electronic record itself. The implications of this difference are addressed in the Discussion.

Data Access and Extraction

Access was obtained through PhysioNet by creating an account and agreeing to the terms of use, completing the credentialed-user application, completing the required CITI program training ("Data or Specimens Only Research") and uploading the completion report, awaiting credential verification, and signing the data usage agreement.

Individual files were downloaded rather than the complete 9.8 GB archive, in order to work within available storage and

memory. From MIMIC-IV (v3.1): admissions, d_labitems, diagnoses_icd, labevents, and patients. From MIMIC-IV-Note: discharge and discharge_detail. This reduced storage requirements by more than 61%.

The process of ETL (Extract, Transform, and Load) was then followed using Python, with Pandas for data manipulation and NumPy for array operations.

Cohort Definition and Inclusion Criteria

Patients were identified by ICD diagnosis codes across both coding systems present in MIMIC-IV. Both ICD-9 and ICD-10 codes were used, so that patients recorded under the earlier coding system were retained. The complete code list and the ICD-9 to ICD-10 correspondence are given in Table 1.

Table 1 ICD-9 and ICD-10 Codes Used For Cohort Definition

Group	ICD-10	ICD-9	Role
Crohn's disease (label 1)	K50.x	555.x	Inclusion
Irritable bowel syndrome (label 0)	K58.x	564.1	Inclusion
Ulcerative colitis	K51.x	556.x	Exclusion
Other non-infective gastroenteritis and colitis	K52.x	558.x	Exclusion
Intestinal infectious disease	A00-A09	001-009	Exclusion
Coeliac disease	K90.0	579.0	Reported, not excluded

Labels were assigned at the patient level rather than the encounter level. A patient with at least one Crohn's disease code and no IBS code was assigned label 1 (Crohn's disease), and a patient with at least one IBS code and no Crohn's disease code was assigned label 0 (IBS). Patients carrying codes for both conditions were excluded rather than assigned to either class, since their true classification cannot be established from coding alone.

Patients coded for ulcerative colitis, other non-infective gastroenteritis and colitis, or intestinal infectious disease were excluded, as these conditions confound the distinction under study. Coeliac disease was recorded and reported but not excluded, as its symptom profile is separable and its exclusion would not materially alter the cohort.

The derivation of the analytic cohort is given in Table 2.

The index admission for each patient was defined as the earliest admission at which a qualifying diagnosis code was recorded. All laboratory and clinical note features were restricted to this admission, so that the model draws on information contemporaneous with the diagnostic episode rather than information generated during subsequent care.

Regarding the reference standard, the labels used in this study are administrative ICD codes. MIMIC-IV does not

Table 2 Cohort Derivation From MIMIC-IV (V3.1)

Step	n
Unique patients in MIMIC-IV (v3.1)	364,627
Patients with any Crohn's disease code	2,422
Patients with any IBS code	3,606
Excluded: coded for both conditions	134
Excluded: ulcerative colitis	203
Excluded: other non-infective colitis	432
Excluded: intestinal infectious disease	462
Analytic cohort	4,663
Crohn's disease	1,772
IBS	2,891
Majority-class baseline accuracy	62.0%
Coeliac disease within analytic cohort (retained)	56

record whether a given diagnosis was established endoscopically, histologically, or clinically, and it contains no field confirming that a patient labeled IBS underwent negative endoscopy or normal imaging to exclude organic disease. Given that approximately 10% of IBD patients are misdiagnosed as IBS even in high-income settings, a proportion of the labels used here is expected to be incorrect. The models are therefore trained to reproduce coded diagnoses, which places an upper bound on measurable accuracy independent of modeling choices. This bound applies equally to all configurations compared in this study, so the relative comparisons between models remain valid.

Laboratory Feature Construction

Laboratory events were loaded from `labevents.csv.gz` in a chunk-based format due to the large file size, with a chunk size of 2,000,000 rows, and filtered to the cohort's index admissions. Each measurement was linked to its textual test name through `d.labitems.csv.gz`. For each patient and each test, the mean value across the index admission was taken, matching the aggregation used in the original analysis.

A laboratory test was retained if a value was present for at least 50% of the cohort. 391 distinct tests were present before this filter and 23 after it. Missing values were imputed with the cohort median and accompanied by explicit binary missingness indicators, since the absence of an ordered test carries clinical information and filling it with zero would encode it as a physiologically implausible measurement.

A basic panel subset was defined separately, restricted to tests obtainable from a complete blood count, a basic metabolic panel, liver function tests, albumin, C-reactive protein, and erythrocyte sedimentation rate. This subset, containing 17 tests, represents laboratory capability plausibly available in a district hospital in an LLMIC, and it was evaluated as a separate configuration so that performance under realistic

input constraints could be measured directly rather than assumed.

Clinical Note Variants

Discharge summaries are authored at the conclusion of an admission, after a diagnosis has been established, and they commonly state that diagnosis explicitly. A model trained on the complete note can therefore achieve high apparent accuracy by detecting the recorded diagnosis rather than by recognizing a clinical pattern. Two text variants were constructed so that this effect could be measured rather than assumed absent.

The raw variant consists of the complete set of discharge summaries for the index admission, concatenated per patient. This reproduces the text input used in the original analysis.

The leakage-controlled variant was constructed using a section allowlist. Notes were parsed into their section headers, and only sections describing the patient's presentation and diagnostic workup were retained: chief complaint, history of present illness, review of systems, physical examination, pertinent results, family history, and social history. Sections authored at or after the point of diagnosis were removed in full: discharge diagnosis, final and primary diagnosis, brief hospital course, discharge medications, medications on admission, discharge instructions, discharge condition, discharge disposition, follow-up instructions, past medical history, and major surgical or invasive procedure. Past medical history was removed because it records prior diagnosis of the condition being predicted, and medication lists were removed because disease-specific agents identify the diagnosis directly.

Within the retained text, residual diagnostic terms were redacted, covering variants of Crohn's, regional enteritis, terminal ileitis, inflammatory bowel disease, IBD, irritable bowel syndrome, IBS, spastic colon, and ulcerative colitis. A list of 62 IBD-specific and IBS-specific medication names was redacted under both generic and brand forms, spanning biologics, small-molecule agents, aminosalicylates, disease-specific corticosteroids, immunomodulators, and IBS-specific agents.

An allowlist was selected in preference to term deletion because retention of a defined set of pre-diagnosis sections can be verified, whereas the removal of every diagnostic phrase from a full document cannot. Clinical findings such as granulomas, skip lesions, cobblestoning, fistulae, and terminal ileal inflammation were deliberately retained, as these constitute legitimate diagnostic signal rather than recorded labels.

Text Vectorization

Vectorization was essential in the development of the ML model, as natural language processing models operate through numerical representations called embeddings rather than

string-form text. ClinicalBERT, a BERT model fine-tuned on clinical text data and medical notes, was chosen over TF-IDF as it supports named entity recognition (NER) and medical natural language inference (MedNLI), giving it the ability to classify medical entities such as diseases, medications, and lab values and to determine the relationship between clinical statements, whereas TF-IDF tracks word frequency without deeper analysis. TF-IDF was nonetheless retained as a comparison baseline, so that the benefit of the richer representation could be quantified rather than assumed.

Clinical notes were embedded using the checkpoint emilyalsentzer/Bio_ClinicalBERT, producing 768-dimensional representations. Documents were tokenized without special tokens, divided into non-overlapping windows of 510 tokens, and capped at 16 windows per patient. Each window was passed through the model and pooled by attention-masked mean pooling across its tokens, and the resulting window embeddings were combined into a single document embedding by a mean weighted by the number of real tokens in each window. This windowed approach was adopted so that the full clinical record is represented rather than only its opening passage, which a single 512-token truncation would otherwise impose.

The numerical embeddings were then combined with the structured laboratory features described above to form the complete feature matrix.

Model Development and Evaluation Protocol

The cohort was partitioned once, at the patient level, with stratification by class, applying the split ratio used in medical applications: 70% for training, 15% for validation and tuning, and 15% for final evaluation. Because all features are aggregated to one row per patient, no patient appears in more than one partition, and multiple admissions belonging to the same patient cannot be distributed across the split.

The test partition was created before any modeling decision was made and was not consulted during architecture selection, hyperparameter search, decision-threshold selection, or early stopping, all of which were carried out using the training and validation partitions only. Each configuration was evaluated on the test partition exactly once, and the reported figure is that single evaluation. This protocol yields a valid held-out estimate and is applied uniformly across every model compared.

The neural model was defined as a sequential architecture with three hidden layers of decreasing width followed by a single sigmoid output unit. After initial testing, class imbalance was addressed using balanced class weights computed on the training partition through Scikit-learn's `compute_class_weight` function. Further testing identified overfitting, attributable to the clinical similarity between Crohn's disease and IBS, the

finite quantity of training data available for a complex medical distinction, and overtraining. Four techniques were applied to reduce overfitting: L2 regularization (ridge regression), dropout, Leaky ReLU activation, and early stopping. L2 regularization was applied with the following cost equation:

$$\text{Loss} + \lambda * (1/2) * \sum (w_i)^2$$

The “loss” represents the original loss function of the model, “ λ ” (lambda) is a hyperparameter that controls the strength of the regularization, and “ $\sum (w_i)^2$ ” is the sum of the squares of all the model's weights. Early stopping was applied with a patience value of 10, meaning that if validation loss had not improved over the previous 10 epochs, training would cease, with a maximum of 150 epochs. The optimizer was Adam and the loss function was binary cross-entropy, with learning rate and batch size selected during hyperparameter search.

Bayesian optimization was implemented for hyperparameter tuning using a Tree-structured Parzen Estimator, which constructs a probabilistic model of the objective and selects successive configurations from it. The search maximized validation ROC AUC. Each search ran for 40 trials for the non-neural models and 12 trials for the neural model, where a trial denotes one complete hyperparameter configuration trained to convergence under the early-stopping criterion above, which is distinct from a training epoch.

To establish whether the complexity of the pipeline is warranted, the following models were evaluated under an identical protocol on identical partitions: a majority-class baseline that always predicts the more frequent class; logistic regression with standardized features; random forest; histogram-based gradient boosting; TF-IDF with a linear classifier applied to both note variants; and the neural network described above. Six feature sets were evaluated across these models: basic-panel laboratory values; all retained laboratory values; raw note embeddings; leakage-controlled note embeddings; laboratory values combined with raw notes; and laboratory values combined with leakage-controlled notes. All configurations were evaluated on the same patients, so that the ablation constitutes a controlled comparison.

The decision threshold was selected on the validation partition by maximizing Youden's J statistic and then applied unchanged to the test partition. Reported metrics comprise accuracy, balanced accuracy, macro F1-score, ROC AUC, average precision, sensitivity, specificity, positive predictive value, negative predictive value, Brier score, and expected calibration error, together with the full confusion matrix. Ninety-five percent confidence intervals were obtained by percentile bootstrap with 2,000 resamples, and differences in ROC AUC between configurations were assessed by paired bootstrap on the same resamples rather than by comparison of point estimates.

Calibration was assessed through reliability curves alongside the Brier score and expected calibration error. Clinical

utility was assessed through decision-curve analysis across threshold probabilities from 0.05 to 0.95, comparing the net benefit of the model against strategies of investigating all patients and investigating none.

Model behavior was examined through permutation importance computed on the test partition over 2,000 permutation rounds, and through SHAP values where the model class supports them. Named laboratory features were used directly rather than principal components for this purpose, so that the resulting attributions correspond to interpretable clinical quantities. For completeness, the explained variance of a principal component analysis of the laboratory block is also reported.

Misclassified patients were characterized by age, sex, length of stay, number of admissions, and predicted probability, and were separated into false negatives, in which Crohn’s disease was classified as IBS, and false positives, in which IBS was classified as Crohn’s disease. Performance was additionally computed within subgroups defined by sex, age band, race, insurance type, and admission type, for subgroups containing at least 20 test patients.

Through all sections of the research, adequate graphs were collected through the Matplotlib package. Analyses were performed in Python using Pandas, NumPy, Scikit-learn, PyTorch, Transformers, and Optuna. All random seeds were fixed at 42, and all post-processed data, analytical results, and ML models were stored in dedicated research repositories.

Ethics Approval And Data Use

This study used MIMIC-IV, which contains fully de-identified patient data and was approved by the institutional review boards of MIT and BIDMC. As all data were de-identified and available to credentialed users under a data use agreement, this secondary analysis did not constitute human subjects research as defined by United States federal regulations, and additional institutional review board approval was not required. Obligations arising from the prospective deployment of a clinical tool, which this determination does not address, are considered in the Discussion.

Results

Cohort Characteristics

The analytic cohort comprised 4,663 patients, of whom 1,772 (38.0%) had Crohn’s disease and 2,891 (62.0%) had IBS. Index-admission laboratory values were available for 4,187 patients and discharge notes for 3,133. The 2,998 patients holding both formed the modelling set, so that every configuration was evaluated on identical patients. Characteristics are given in Table 3.

Table 3 Characteristics of the Analytic Cohort

Characteristic	Crohn’s disease (n = 1,772)	IBS (n = 2,891)
Median age at index admission, years (IQR)	52 (35-66)	61 (45-73)
Female, %	54.0	77.7
Median admissions per patient	2	2
Median length of stay, days	3.3	2.9

Of 391 distinct laboratory tests recorded for the cohort, 23 were present for at least 50% of patients and were retained. Seventeen of these 23 tests fall within the basic panel definition, indicating that the laboratory information available for this cohort consists predominantly of routine tests rather than specialized assays.

Following the patient-level stratified split, the training partition contained 2,098 patients, the validation partition 450, and the test partition 450. The balanced class weights computed on the training partition were 0.856 for IBS and 1.202 for Crohn’s disease.

Label Leakage Audit

Before scrubbing, 2,898 of 3,133 patient records (92.5%) contained an explicit diagnostic term or a disease-specific medication name within their discharge summary. After section filtering and redaction, this figure fell to 48 records (1.5%), a reduction of 98.3%. Median record length fell from 9,346 to 2,311 characters, and 6 patients had no retained text.

The great majority of discharge summaries in this cohort therefore state the diagnosis directly, confirming that any model trained on unfiltered summaries has access to the outcome it is asked to predict.

Model Performance

Performance for all configurations is given in Table 4. The primary configuration was selected by validation ROC AUC among the leakage-controlled configurations and evaluated once on the locked test partition.

The primary leakage-controlled configuration achieved an accuracy of 76.67% (95% CI 72.67-80.67), a balanced accuracy of 75.25% (95% CI 71.18-79.15), a macro F1-score of 75.58% (95% CI 71.43-79.48), and an AUC of 83.87% (95% CI 79.94-87.61) at a decision threshold of 0.509 selected on the validation partition. Sensitivity for Crohn’s disease was 66.84% (95% CI 60.10-73.40) and specificity was 83.65% (95% CI 79.20-88.17), with a positive predictive value of 74.40% and a negative predictive value of 78.01%. The

Table 4 Test-Set Performance Across Models and Feature Sets

Configuration	Accuracy % (95% CI)	Macro F1 %	AUC % (95% CI)	Sens. %	Spec. %
Majority-class baseline	58.4 (53.8-62.7)	36.9	50.0	0.0	100.0
Basic labs, logistic regression	62.7 (58.5-74.1)	61.2	65.8 (60.9-71.3)	74.1	61.7
Basic labs, random forest	61.3 (56.4-65.8)	60.9	66.3 (60.9-71.3)	60.4	62.0
All labs, random forest	60.0 (55.3-64.7)	60.0	66.3 (60.8-71.4)	63.1	58.2
ClinicalBERT, leakage-controlled	70.0 (66.0-74.2)	69.8	79.3 (75.0-83.6)	73.8	67.3
Labs + ClinicalBERT, leakage-controlled	74.0 (70.0-78.0)	73.1	79.4 (75.1-83.6)	66.8	79.1
TF-IDF, leakage-controlled (primary)	76.7 (72.7-80.7)	75.6	83.9 (79.9-87.6)	66.8	83.7
ClinicalBERT, unredacted notes	85.8 (82.4-88.7)	84.9	90.0 (86.7-92.9)	74.3	93.9
TF-IDF, unredacted notes	86.9 (83.6-89.8)	86.4	92.6 (89.9-94.9)	81.8	90.5

confusion matrix comprised 220 true negatives, 43 false positives, 62 false negatives, and 125 true positives.

These figures exceed those reported for the previous configuration of this model, which achieved an accuracy of 74.41%, an F1-score of 73.40%, and an AUC of 79.56%. The present figures are obtained under a locked test set, using only text authored before diagnostic confirmation, and are therefore both higher and more conservatively derived.

Effect of Diagnostic Text Removal

Comparing identical model classes trained on unredacted and redacted text isolates the contribution of recorded diagnostic language. Using TF-IDF features, AUC fell from 92.63% to 83.87%, a difference of 0.0874 (95% CI 0.0626-0.1143, $p < 0.001$). Using ClinicalBERT embeddings, AUC fell from 90.01% to 79.27%, a difference of 0.1074 (95% CI 0.0679-0.1482, $p < 0.001$). With laboratory values added to ClinicalBERT embeddings, AUC fell from 89.58% to 79.44%, a difference of 0.1014 (95% CI 0.0623-0.1387, $p < 0.001$).

Approximately 0.09 to 0.11 AUC in this task is therefore attributable to diagnostic text recorded in the discharge summary. The remaining performance, an AUC of 83.87%, is obtained from presenting symptoms, examination findings, and pertinent results alone, and represents diagnostic signal rather than recorded outcome.

Comparison with Baseline and Simpler Models

The majority-class baseline achieved 58.44% accuracy with an AUC of 0.500 by construction. The primary configuration exceeded this baseline by 18.2 percentage points of accuracy and by 0.3385 AUC (95% CI 0.2985-0.3765, $p < 0.001$).

Models restricted to laboratory values achieved AUCs between 60.29% and 66.26%, with the basic 17-test panel performing equivalently to the full 23-test panel. The primary configuration exceeded the strongest laboratory model by 0.1754 AUC (95% CI 0.1212-0.2289, $p < 0.001$), indicating that most of the recoverable signal in this cohort resides in the clinical narrative rather than in routine laboratory values.

Among leakage-controlled text configurations, TF-IDF with a linear classifier outperformed ClinicalBERT embeddings by 0.0463 AUC (95% CI 0.0153-0.0793, $p = 0.005$) and outperformed the combined laboratory and ClinicalBERT configuration by 0.0443 AUC (95% CI 0.0115-0.0776, $p = 0.008$). The deep neural network did not outperform logistic regression on any feature set. The best-performing configuration in this study is therefore also the least computationally demanding, a finding with direct implications for deployment feasibility that are considered in the Discussion.

Calibration and Clinical Utility

The primary configuration produced a Brier score of 0.1797 (95% CI 0.1698-0.1897) and an expected calibration error of 0.1445. Reliability analysis showed that predicted probabilities overstate risk in the lower range, with a mean predicted probability of 0.354 against an observed frequency of 0.147 in the 0.3 to 0.4 bin, and understate risk in the upper range, with a mean predicted probability of 0.749 against an observed frequency of 1.000 in the 0.7 to 0.8 bin. The ranking of patients is therefore substantially more reliable than the absolute probabilities, and recalibration would be required before predicted probabilities were used as risk estimates.

Decision-curve analysis showed that the model delivers positive net benefit across threshold probabilities from 0.30 to 0.80, and exceeds both the investigate-all and investigate-none strategies throughout this range. At a threshold probability of 0.50, the net benefit of the model was 0.187, compared with negative 0.169 for investigating all patients and 0 for investigating none. The model therefore adds clinical value across the range of thresholds plausible for a triage decision.

Class-Weight Ablation

Balanced class weights of 0.856 for IBS and 1.202 for Crohn's disease were applied during training. Removing them reduced AUC from 76.24% to 73.90% and accuracy from 71.78% to 64.89% in the matched configuration, confirming that the

weighting improved performance rather than being applied without effect.

Error Analysis

Of 450 test patients, 345 were classified correctly and 105 were misclassified, comprising 62 false negatives, in which Crohn’s disease was classified as IBS, and 43 false positives.

No misclassification was made at high confidence. Every error occurred at a predicted probability between 0.2 and 0.8, meaning the model expressed uncertainty on the cases it got wrong rather than asserting an incorrect diagnosis confidently. For a triage application this is the preferable failure mode, since low-confidence predictions can be routed to clinical review.

The two error types differed systematically. Patients whose Crohn’s disease was missed were older than correctly classified patients (median 63.5 against 57.0 years), more frequently female (83.9% against 68.4%), and had longer admissions (median 4.6 against 3.5 days). Patients incorrectly assigned to Crohn’s disease were predominantly male. Missed Crohn’s disease therefore concentrates in older female patients with longer admissions, a group in which symptoms are more readily attributed to functional causes, mirroring the clinical misdiagnosis pattern this study addresses.

Subgroup Performance

Performance across demographic and care-setting subgroups is given in Table 5.

Table 5 Primary Model Performance Within Subgroups

Subgroup	n	AUC	Accuracy	Balanced accuracy
Female	305	0.806	0.774	0.699
Male	145	0.855	0.752	0.718
Age under 30	53	0.991	0.906	0.922
Age 30-44	77	0.869	0.753	0.751
Age 45-59	113	0.846	0.779	0.778
Age 60-74	113	0.766	0.717	0.690
Age 75 and over	94	0.751	0.745	0.622
White	381	0.836	0.769	0.752
Black	24	0.896	0.750	0.733
Other or unknown	24	0.895	0.833	0.818
Private insurance	177	0.890	0.808	0.808
Medicare	202	0.797	0.752	0.698
Medicaid	57	0.718	0.632	0.607

Performance was strongest in younger patients, reaching an AUC of 0.991 under age 30, and declined to 0.751 in patients aged 75 and over. This gradient is consistent with the clinical picture, as Crohn’s disease typically presents in early adulthood and the differential in older patients is wider. Perfor-

mance was consistent across racial groups. Performance was lower among Medicaid-insured patients (AUC 0.718) than privately insured patients (0.890), a disparity that likely reflects differences in the completeness of the documented record rather than in the underlying disease, and is reported here so that it is available for scrutiny.

Discussion

The ML model developed in this study, which differentiates Crohn’s disease and IBS using only clinical information recorded before diagnostic confirmation, achieved an accuracy of 76.67%, a macro F1-score of 75.58%, and an AUC of 83.87% (Table 4). These figures exceed those obtained by the previous configuration of this model, which achieved 74.41% accuracy and an AUC of 79.56%, and they were obtained under a stricter protocol using less information. The model exceeds a majority-class baseline by 18.2 percentage points of accuracy and by 0.34 AUC, and it delivers positive net benefit across the full range of threshold probabilities plausible for a triage decision.

Effect of Diagnostic Text on Result Validity

The central methodological finding of this study is that 92.5% of the discharge summaries in this cohort contained the diagnosis or a disease-specific medication in plain text, and that removing this information costs between 0.09 and 0.11 AUC depending on the representation used. Any model in this domain trained on complete discharge summaries is therefore reading a substantial part of its answer from the record.

The consequence for this study is favorable. After removing that information, an AUC of 83.87% remains, obtained from presenting symptoms, examination findings, and pertinent results alone. This is a measurement of diagnostic signal rather than of documentation practice, and it is the figure on which the remainder of this discussion rests. Quantifying this separation is uncommon in clinical natural language processing, and the ability to state precisely how much performance derives from recorded outcomes rather than from clinical evidence strengthens the interpretation of the result considerably.

Model Complexity and Performance

Among leakage-controlled configurations, TF-IDF with logistic regression outperformed ClinicalBERT embeddings by 0.046 AUC and outperformed every neural configuration tested. Two factors plausibly contribute. The redacted records are short, with a median of 2,311 characters, and consist largely of symptom and finding vocabulary, which a sparse lexical representation captures efficiently. With 2,098 training patients, a dense 768-dimensional representation feeding

a multi-layer network also carries more capacity than the data can constrain.

This result carries direct practical weight. A TF-IDF and logistic regression model occupies a few megabytes, trains in seconds, produces a prediction in under a millisecond on an ordinary processor, and requires no accelerator, no network connection, and no specialized runtime. The configuration that performed best in this study is therefore also the configuration that is most straightforward to deploy, which is an unusual and welcome alignment between statistical performance and practical feasibility.

Cost and Accessibility

The clinical inputs required by the primary model are a written history, the presenting complaint, examination findings, and available results. These are produced in the ordinary course of a consultation and do not require endoscopy, biologic assay, or specialist review. This matters directly given that 52.4% of Crohn's disease patients in Asia and 60.9% in Africa cannot afford diagnostic testing,¹³ that Eastern Africa holds approximately 0.12 endoscopists and 0.09 colonoscopes per 100,000 inhabitants,¹⁷ and that gastroenterologist density across surveyed LLMICs falls between 0.005 and 0.02 per 100,000 inhabitants.

The primary model requires no graphics hardware for training or inference. The ClinicalBERT configurations require accelerated hardware and a model download of several hundred megabytes, and performed less well. The binding constraint on a tool of this design is therefore neither compute nor laboratory infrastructure, but the existence of a written clinical record the model can read, a substantially lower requirement than the endoscopic pathway it would triage toward.

Faecal calprotectin distinguishes inflammatory from functional bowel disease without endoscopy, and abdominal ultrasound is constrained by operator training rather than equipment cost. This study does not compare against either, and a model of this kind would most plausibly function alongside them, identifying which patients warrant the test rather than replacing it.

Implementation and Ethical Considerations

A tool that influences whether a patient is referred for endoscopy participates in a clinical decision. Responsibility for erroneous predictions must be assigned before any deployment, and rests with the deploying institution rather than with the model. Deployment would require local validation, monitoring for performance drift, a defined retraining process, and a named maintaining party. Patient consent for secondary use of records, local data governance approval, and review by health authorities in any deploying country are prerequisites.

The populations this work is directed toward were not consulted in its design, and their involvement would be necessary in any subsequent development.

The error analysis reported above is relevant to this discussion. No misclassification in this study was made at high confidence, which means that a deployment routing low-confidence predictions to clinical review would capture every error the model made on the test partition. A triage design that defers rather than decides on uncertain cases is therefore supported by the observed behavior of this model.

Interpretability

The primary model is a linear classifier over lexical features, so each prediction decomposes into the contribution of individual terms in the record, and a clinician can inspect which elements of a history drove a given score. The transformer configurations do not offer this, nor does a principal-component representation of laboratory values, which is why named laboratory features were used throughout. Where model behavior informs a referral decision, the ability to inspect its basis is a requirement rather than an enhancement, and the best-performing model here provides it.

Limitations

The reference standard is administrative ICD coding rather than endoscopic or histological confirmation, so the models reproduce coded diagnoses, a proportion of which is expected to be incorrect. This constraint applies identically to every configuration compared, so the relative comparisons remain valid, but absolute figures should be understood as measured against coding rather than verified truth.

No data from any LLMIC were used at any stage. The performance reported here characterizes this cohort and establishes that the diagnostic signal exists in routinely recorded clinical text. It does not establish performance in any other health system, and external validation in the target setting is the necessary next step rather than an optional extension. The population studied consists of hospitalized patients, who are more severely affected than the outpatient populations in which this distinction is usually drawn, which represents a spectrum difference relative to the eventual use case. The test partition of 450 patients bounds the precision of all estimates, and confidence intervals are reported throughout so that this precision is visible. Features were restricted to the index admission, which excludes longitudinal information a clinician would ordinarily use. Redaction reduced records containing diagnostic terms by 98.3% but cannot eliminate diagnostic paraphrase, as a description such as granulomatous inflammation of the terminal ileum carries diagnostic weight without containing any redacted term. MIMIC-IV covers admissions

between 2008 and 2019, and coding and treatment practices have developed since.

Predicted probabilities are moderately miscalibrated, as reported above, and recalibration would be required before they were interpreted as risk estimates. Model ranking, which is what a triage application depends upon, is considerably more reliable.

Comparison with The Previous Analysis

Several figures differ from those obtained previously for methodological reasons. The cohort is larger, at 4,663 patients against 3,493, and its class balance differs, at 62.0% IBS, following patient-level labelling with exclusion of patients coded for both conditions, inclusion of ICD-9 codes alongside ICD-10, and exclusion of confounding diagnoses. The denominator is corrected to 364,627 patients. Test performance is reported once per configuration under a locked test set rather than selected across candidates. Every headline metric is nonetheless higher than previously reported, obtained from strictly less information.

Conclusion

This study developed and analyzed ML models that classify the distinction between Crohn's disease and IBS using routinely recorded clinical information, drawing on a cohort of 4,663 patients from MIMIC-IV, of whom 2,998 held both laboratory and clinical note data. Using only text authored before diagnostic confirmation, and evaluated once on a test partition locked before any modelling decision, the model achieved an AUC of 83.87% (95% CI 79.94-87.61), an accuracy of 76.67% (95% CI 72.67-80.67), and a macro F1-score of 75.58%, against a majority-class baseline of 58.44%. Decision-curve analysis showed positive net benefit across threshold probabilities from 0.30 to 0.80, and no misclassification was made at high confidence.

This study further establishes that 92.5% of the discharge summaries in this cohort state the diagnosis in plain text, and that this recorded information accounts for between 0.09 and 0.11 AUC. Separating that contribution from genuine diagnostic signal is uncommon in clinical natural language processing, and it allows the performance reported here to be interpreted as a measurement of clinical evidence rather than of documentation practice.

The configuration that performed best is also the most economical. It requires no accelerated hardware, no laboratory panel, and no specialist assessment, drawing only on the written history, examination, and available results that a consultation ordinarily produces. In settings where roughly half to two-thirds of patients cannot afford diagnostic testing and a single gastroenterologist may serve millions of people, the

value of a method whose inputs are already generated and whose compute requirements are negligible is considerable.

These findings are preliminary and derive from a single retrospective dataset with administrative diagnosis codes as the reference standard. They do not establish that the model can diagnose Crohn's disease, and the model is not a substitute for endoscopic evaluation. What they do establish is that the information needed to separate these two conditions is present in ordinary clinical documentation, that it can be recovered without specialized infrastructure, and that a tool built on it identifies patients warranting further investigation with measurable clinical benefit. Given that patients with Crohn's disease in low- and lower-middle-income countries currently wait roughly twice as long for a diagnosis as those in high-income settings, establishing that this signal exists and is cheaply recoverable is a meaningful step toward closing a gap that existing diagnostic infrastructure has not closed.

Ethical Considerations

This study utilized data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. The MIMIC-IV database contains fully de-identified patient data and was approved by the Institutional Review Boards of the Massachusetts Institute of Technology (MIT) and Beth Israel Deaconess Medical Center (BIDMC). As all data were de-identified and publicly available to credentialed users under a data use agreement, this study did not involve human subjects research as defined by U.S. federal regulations, and additional Institutional Review Board (IRB) approval was not required. This determination addresses the secondary analysis conducted here; the obligations arising from prospective clinical deployment are addressed in the Discussion.

Conflicts of Interest

The author declares no conflicts of interest.

Author Contributions

The author gathered and extracted the data, developed predictive ML models, conducted the study, interpreted the results, and wrote the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The data used in this study were obtained from the dataset MIMIC-IV (V3.1) and MIMIC-IV-Note, available publicly through PhysioNet. Credentialed registration and CITI training are prerequisites to obtaining access to the dataset. No new data were generated through this study. Analysis code, including cohort construction, leakage control, and evaluation scripts, is available from the author on request.

References

- 1 D. Warner. IBS vs. Crohn's disease: What to know. Medical News Today. 2024. <https://www.medicalnewstoday.com/articles/ibs-vs-crohns>.
- 2 E. Cross, B. Saunders, A. D. Farmer, J. A. Prior. Diagnostic delay in adult inflammatory bowel disease: A systematic review. *Indian Journal of Gastroenterology*. Vol. 42, pg. 40–52, 2023. <https://doi.org/10.1007/s12664-022-01303-x>.
- 3 New study finds delayed diagnosis of inflammatory bowel disease is linked to worse clinical outcomes. Crohn's & Colitis UK. 2023. <https://crohnsandcolitis.org.uk/news-stories/news-items/new-study-shows-delayed-diagnosis-of-inflammatory-bowel-disease-is-linked-to-worse-clinical-outcomes>.
- 4 Cleveland Clinic. Crohn's disease. Cleveland Clinic. 2023. <https://my.clevelandclinic.org/health/diseases/9357-crohns-disease>.
- 5 Mayo Clinic Staff. Irritable bowel syndrome: Symptoms and causes. Mayo Clinic. 2024. <https://www.mayoclinic.org/diseases-conditions/irritable-bowel-syndrome/symptoms-causes/syc-20360016>.
- 6 W. D. Chey, J. Kurlander, S. Eswaran. Irritable bowel syndrome: A clinical review. *JAMA*. Vol. 313, pg. 949, 2015. <https://doi.org/10.1001/jama.2015.0954>.
- 7 S. Berg. What doctors wish patients knew about irritable bowel syndrome. American Medical Association. 2023. <https://www.ama-assn.org/delivering-care/prevention-wellness/what-doctors-wish-patients-knew-about-irritable-bowel-syndrome>.
- 8 ACE Clinic. Recognizing and dealing with a Crohn's flare-up. Arizona Colorectal Experts. 2024. <https://aceclinic.org/recognizing-and-dealing-with-a-crohns-flare-up/>.
- 9 Colorectal Clinic of Tampa Bay. How to calm an IBS flare-up. 2021. <https://www.tampacolorectal.com/blog/how-to-calm-an-ibs-flare-up>.
- 10 S. J. Halpin, A. C. Ford. Prevalence of symptoms meeting criteria for irritable bowel syndrome in inflammatory bowel disease: Systematic review and meta-analysis. *American Journal of Gastroenterology*. Vol. 107, pg. 1474–1482, 2012. <https://doi.org/10.1038/ajg.2012.260>.
- 11 K. M. Fairbrass, S. J. Costantino, D. J. Gracie, A. C. Ford. Prevalence of irritable bowel syndrome-type symptoms in patients with inflammatory bowel disease in remission: A systematic review and meta-analysis. *The Lancet Gastroenterology & Hepatology*. Vol. 5, pg. 1053–1062, 2020. [https://doi.org/10.1016/S2468-1253\(20\)30300-9](https://doi.org/10.1016/S2468-1253(20)30300-9).
- 12 N. Jayasooriya, S. Baillie, J. Blackwell, et al. Systematic review with meta-analysis: Time to diagnosis and the impact of delayed diagnosis on clinical outcomes in inflammatory bowel disease. *Alimentary Pharmacology & Therapeutics*. Vol. 57, pg. 635–652, 2023. <https://doi.org/10.1111/apt.17370>.
- 13 R. Rajbhandari, S. Blakemore, N. Gupta, et al. Crohn's disease among the poorest billion: Burden of Crohn's disease in low- and lower-middle-income countries. *Digestive Diseases and Sciences*. Vol. 68, pg. 1226–1236, 2022. <https://doi.org/10.1007/s10620-022-07675-6>.
- 14 T. R. Card, J. Siffledeen, K. M. Fleming. Are IBD patients more likely to have a prior diagnosis of irritable bowel syndrome? Report of a case-control study in the General Practice Research Database. *United European Gastroenterology Journal*. Vol. 2, pg. 505–512, 2014. <https://doi.org/10.1177/2050640614554217>.
- 15 C. K. Porter, B. D. Cash, M. Pimentel, A. Akinseye, M. S. Riddle. Risk of inflammatory bowel disease following a diagnosis of irritable bowel syndrome. *BMC Gastroenterology*. Vol. 12, pg. 55, 2012. <https://doi.org/10.1186/1471-230X-12-55>.
- 16 R. Rajbhandari, S. Blakemore, N. Gupta, et al. Crohn's disease in low and lower-middle income countries: A scoping review. *World Journal of Gastroenterology*. Vol. 26, pg. 6891–6908, 2020. <https://doi.org/10.3748/wjg.v26.i43.6891>.
- 17 A. Gaur, H. R. Bharadwaj, P. Dalal, M. Ahmed. Recent progress and future directions for expanding gastrointestinal endoscopy in low- and middle-income African nations. *JGH Open*. Vol. 8, 2024. <https://doi.org/10.1002/jgh3.70026>.
- 18 Sharing my story is a superpower: The doctor living without a cure in Ethiopia. *The Guardian*. 2022. <https://www.theguardian.com/global-development/2022/oct/27/doctor-fasika-teferra-superpower-hope-crohns-ethiopia-acc>.
- 19 C. Schönborn, M. Levy, M. D. Jaeger, et al. Unmet health-related needs in patients with Crohn's disease in Belgium: A mixed-methods study. *Archives of Public Health*. Vol. 83, 2025. <https://doi.org/10.1186/s13690-025-01632-1>.
- 20 Crohn's & Colitis Foundation. Current research initiatives. <https://www.crohnscolitisfoundation.org/research/current-research-initiatives>.
- 21 AIPRM. 50+ AI in healthcare statistics 2024. AIPRM. 2024. <https://www.aiprm.com/ai-in-healthcare-statistics/>.
- 22 Grand View Research. Artificial intelligence in healthcare market size report. Grand View Research. 2024. <https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market>.
- 23 Y. Guo, Z. Hao, S. Zhao, J. Gong, F. Yang. Artificial intelligence in health care: Bibliometric analysis. *Journal of Medical Internet Research*. Vol. 22, pg. e18228, 2020. <https://doi.org/10.2196/18228>.
- 24 N. Sahni, G. Stein, R. Zimmel, D. M. Cutler. The potential impact of artificial intelligence on healthcare spending. *National Bureau of Economic Research*. 2023. <https://doi.org/10.3386/w30857>.
- 25 M. Ghezzi. How AI in healthcare could save over 250,000 lives each year and become a \$188 billion market by 2030. *The Journal of mHealth*. 2023. <https://thejournalofmhealth.com/how-ai-in-healthcare-could-save-over-250000-lives-each-year-and-become-a-188-billion-market-by-2030/>.
- 26 N. P. Nsengiyumva, H. Hussain, O. Oxlade, et al. Triage of persons with tuberculosis symptoms using artificial intelligence-based chest radiograph interpretation: A cost-effectiveness analysis. *Open Forum Infectious Diseases*. Vol. 8, 2021. <https://doi.org/10.1093/ofid/ofab567>.
- 27 S. T. Kakileti, H. J. Madhu, L. Krishnan, G. Manjunath, S. Sampangi, H. V. Ramprakash. Observational study to evaluate the clinical efficacy of Thermalix for detecting breast cancer in symptomatic and asymptomatic women. *JCO Global Oncology*. Vol. 6, pg. 1472–1480, 2020. <https://doi.org/10.1200/GO.20.00168>.
- 28 K. Adapa, A. Gupta, S. Singh, et al. A real-world evaluation of an innovative artificial intelligence tool for population-level breast cancer screening. *npj Digital Medicine*. Vol. 8, 2025. <https://doi.org/10.1038/s41746-024-01368-2>.