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.

PRISM - Predictive Reasoning and Interactive Sequential Modeling for Disease Management

Mohammad Assadi Shalmani
University of Wisconsin-Milwaukee
Masoud Khani
University of Wisconsin-Milwaukee
Preetika Sinh
University of California
Qiang Lu
China University of Petroleum
Jake Luo
jakeluo@uwm.edu

University of Wisconsin-Milwaukee

Research Article

Keywords: Clinical Decision Making, Predictive Models, Disease Management, Generative Artificial Intelligence, Multi-agent Systems, Large Language Models

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11322733/v1

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

Additional Declarations: No competing interests reported.

PRISM - Predictive Reasoning and Interactive Sequential Modeling for Disease Management

Mohammad Assadi Shalmani
Health Informatics Program, Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Masoud Khani
Health Informatics Program, Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Preetika Sinh
Division of Gastroenterology and Hepatology, University of California, San Diego, CA, USA

Qiang Lu
Department of Computer Science and Technology, China University of Petroleum, Beijing, China

Jake Luo*
Health Informatics Program, Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

corresponding author: jakeluo@uwm.edu

Background: Longitudinal predictive models can anticipate subsequent diagnoses, but connecting their predictions to clinical interpretation and management remains challenging. We developed PRISM, an agentic framework that combines diagnosis prediction with biomarker analysis, comorbidity assessment, intervention analysis, and knowledge extraction, using inflammatory bowel disease (IBD) as a use case. **Methods:** Using a cohort of 113,729 patients, we developed a decoder-only transformer that predicted diagnoses from structured longitudinal electronic health records. Four specialized agents guided analyses using feature masking, Shapley attribution, medication insertion, and statistical association screening. Evaluation included 318,145 prediction positions from 11,382 held-out patients, with analyses conducted at cohort and individual levels. **Results:** Exact-code Top-1, Top-3, Top-5, and Top-10 accuracy was 15.28%, 28.69%, 36.67%, and 49.04%, respectively. Top-10 accuracy reached 86.52% when any of the next ten diagnoses was accepted. Within outcome-specific baseline Top-1 subgroups, masking conventional CRP, high-sensitivity CRP, and fecal calprotectin reduced mean predicted IBD probability by 3.02, 3.48, and 1.37 percentage points. Masking IBD history reduced predictions of intestinal obstruction, nutritional deficiencies, and inflammatory arthropathies by 4.43, 3.67, and 4.98 points. Prednisone, metronidazole, and vedolizumab insertion reduced mean next-diagnosis nutritional-deficiency probability by 1.89, 1.89, and 1.22 points. Knowledge extraction identified a direct association between IBD and opioid-related disorders and an indirect relationship linking IBD, intestinal obstruction, and subsequent peritonitis or intra-abdominal abscess. **Conclusions:** PRISM integrated longitudinal prediction with clinically guided interpretation, intervention exploration, and cohort-scale association screening. Combining cohort summaries with individual analyses revealed differences obscured by averages and generated hypotheses for clinical investigation.

Keywords: Clinical Decision Making, Predictive Models, Disease Management, Generative Artificial Intelligence, Multi-agent Systems, Large Language Models

1. Introduction

Machine learning and deep learning have been widely applied to clinical risk estimation, and electronic health records are now a common data source for this work [1,2]. These models estimate whether a patient will develop a given condition from variables recorded in that patient's history, but most of these models do not operate on the trajectory itself [1,3]. Their input is a feature vector of fixed length, so a record of arbitrary size must first be reduced to a predetermined set of variables, commonly a few dozen [1]. That reduction is set before training, which leaves the model working from a curated summary rather than from the trajectory the summary stands for, and it discards the order in which events occurred [3,4].

Sequence models address this by treating the record as an ordered series of events rather than a single fixed-length feature vector [5]. Transformers have become the common choice, since self-attention relates events separated by long intervals, and models built this way outperform earlier machine learning approaches on the same prediction tasks [6–8]. In generative form, the architecture extends prediction from a single labeled outcome to the trajectory itself, forecasting subsequent events in a patient timeline [9]. Accuracy alone, however, is not what decides whether a prediction is used at the point of care [10,11]. A model that returns a probability without accounting for it gives the physician no basis on which to accept or question. Explainable artificial intelligence developed in response. Post hoc attribution methods address the opacity, dividing a prediction among the inputs that produced it and ranking those inputs by the size of their contribution, so that a prediction can be examined rather than taken on trust [12,13].

However, attribution-based explanations primarily describe which features contribute to a model's prediction, rather than providing a broader view of the patient's clinical trajectory [14,15]. Such explanations can help clinicians interpret individual predictions and

identify relevant predictors, particularly in diagnostic settings. Once a diagnosis is established, however, predictive models may also support questions related to prevention and treatment, such as which conditions warrant monitoring and how potential interventions may influence future outcomes [16]. Addressing these questions requires more than explaining an individual prediction. It requires models that can analyze longitudinal clinical information and support the evaluation of alternative clinical scenarios. A further challenge is that average treatment effects estimated across patient populations can obscure differences in benefit and harm between individuals [17]. Complementing population-level evidence with patient-specific comparisons of alternative treatment scenarios is therefore important for understanding variation that cohort averages may not capture. Further integration of these capabilities may help extend predictive models beyond diagnosis to support prevention, treatment planning, and longitudinal disease management [18].

These limitations are particularly relevant after diagnosis, when clinicians must monitor different conditions and consider treatment options. We therefore developed PRISM(Predictive Reasoning and Interactive Sequential Modeling), an agentic framework that integrates a longitudinal predictive transformer with specialized large language model agents to analyze predicted disease trajectories and support clinical reasoning. The transformer forecasts subsequent clinical events, while the agents use their understanding of clinical concepts and relationships to examine comorbidities, biomarkers, treatments, and other factors associated with disease progression. This integration enables analyses that extend beyond prediction to provide clinically relevant information for prevention, treatment, and longitudinal disease management. By automating these otherwise labor-intensive analytical steps, the framework enables systematic investigation of longitudinal predictions across large patient cohorts while retaining patient-level analyses.

We evaluated this framework using inflammatory bowel disease as a use case because its clinical course is heterogeneous and extends beyond the gastrointestinal tract, with manifestations involving multiple organ systems and frequent comorbidities [19]. Common comorbidities include intestinal obstruction [20,21], nutritional deficiencies [22], inflammatory arthropathies [19,23], and cardiovascular disease [24]. These diverse and interacting clinical features make IBD a suitable setting for evaluating whether an integrated predictive and agentic approach can identify relevant biomarkers and comorbidities, examine potential treatment-related changes, and extract clinically meaningful relationships from longitudinal records.

This study addresses three questions. First, can an integrated predictive and agentic framework characterize biomarkers and comorbid conditions associated with disease trajectories and identify relationships that are consistent with established clinical knowledge? Second, can clinical knowledge integrated through specialized agents guide the analysis and interpretation of relationships among diseases, biomarkers, and treatments identified from longitudinal patient trajectories? Third, can the framework extract clinically relevant associations from longitudinal records at cohort scale, enabling systematic knowledge discovery that would be difficult to perform through manual review?

2. Methods

Our research approach integrates longitudinal Electronic Health Records(EHR), sequence modeling, and an agentic analytical framework to characterize and analyze IBD progression. The pipeline first extracts and organizes diagnoses, medications, procedures, laboratory results, vital signs, and demographic information from patient records. These data are then temporally structured into patient sequences for model training. A decoder-only transformer generates predictions of subsequent clinical events from each longitudinal sequence. The predicted trajectories are subsequently analyzed by specialized agents that use clinical knowledge and analytical tools to investigate biomarkers, comorbidities, interventions, and other clinically relevant relationships. The framework also provides an interactive interface through which clinicians can query the model and examine the resulting analyses and interpretations.

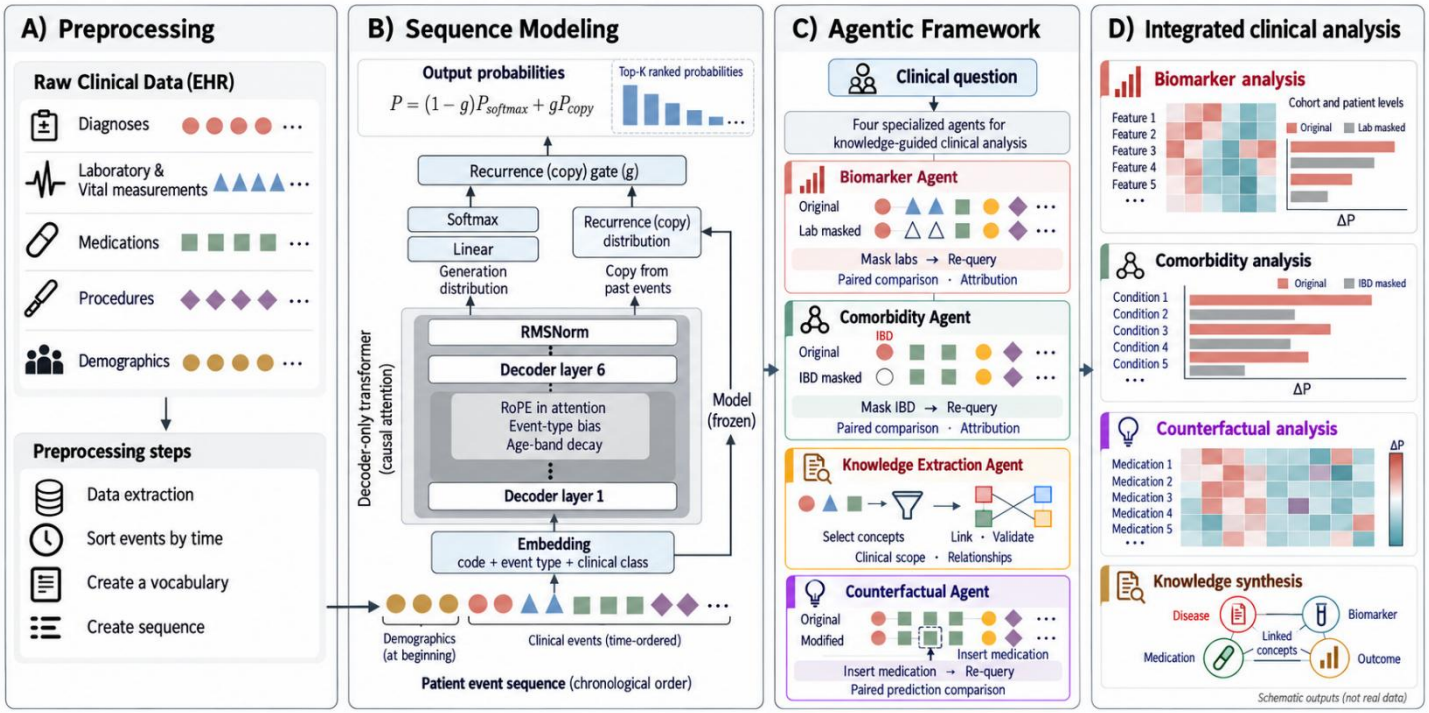


Figure 1. Overview of PRISM for longitudinal diagnosis prediction and agent-guided clinical analysis. (A) Demographics, diagnoses, medications, procedures, vital signs, and laboratory results are extracted and encoded into patient sequences, with demographic tokens preceding time-ordered clinical events. (B) A decoder-only transformer combines a probability distribution over diagnosis codes with a recurrence distribution over previously recorded diagnoses to generate ranked predictions. (C) Four specialized agents: Biomarker, Comorbidity, Counterfactual, and Knowledge Extraction, guide concept selection and interpretation, while computational tools perform feature masking, medication insertion, attribution, and association testing using the frozen model and recorded histories. (D) Outputs summarize biomarker contributions, comorbidity predictions, medication scenarios, and direct and indirect associations at cohort and individual levels. Medication scenarios describe changes in model predictions rather than established treatment effects. Displayed outputs are schematic

Figure 1 illustrates the integration of these components from data preparation through clinical analysis. Rather than using the predictive model solely to generate future events, the framework uses its predictions as a basis for targeted clinical investigation. The agentic layer connects the predicted trajectory with clinically relevant analyses, allowing the same system to examine individual predictions as well as broader relationships across patient cohorts.

2.1 Data Source and Patient Selection

We utilized TriNetX database that combines medical records from 80 hospitals and clinics across the United States, representing more than 40 million patients. This database contains various events including patient's demographics, diagnoses, medications, lab tests, procedures, and vital signs. Our baseline characteristics of the 113,729 IBD patients identified through our systematic search of the TriNetX database. The cohort demonstrates a nearly equal non-exclusive distribution between Crohn's disease (55.9%) and ulcerative colitis (54.8%) patients, with balanced gender representation (51.2% female, 47.0% male). The most prevalent comorbidities reflect common IBD-related complications and symptoms, with noninfective gastroenteritis and colitis being the most frequent additional diagnosis (34.5%), followed by diarrhea (30.0%) and hypertension (25.0%). We identified patients with inflammatory bowel disease (IBD) through a retrospective cohort study utilizing the TriNetX database, covering the period from January 1, 2015, to November 15, 2023.

The age distribution of our IBD cohort is visualized in Figure 2, with the largest number of patients aged 30–35 years. Mean age was 48.3 years (SD, 18.2) showing the typical health patterns seen in people with IBD.

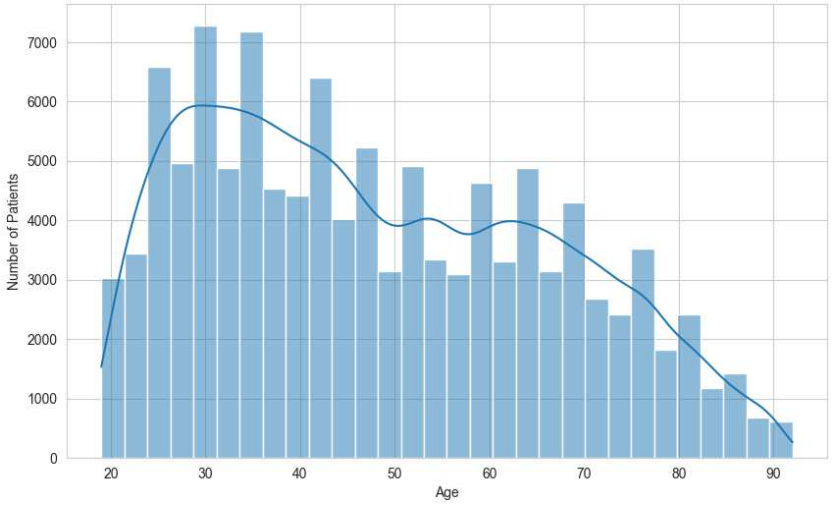


Figure 2. Age distribution of IBD patients in the study cohort

2.2 Predictive Model Architecture and Training

We used a decoder-only transformer with six layers, eight attention heads, a model dimension of 256, and a feed-forward dimension of 1,024. The model contained approximately 5.5 million trainable parameters and each clinical event was represented using a code embedding, an event-type embedding, and a clinical-class embedding. For an event x_t, the initial representation was:

$$h_t^{(0)} = e(x_t) + E^{\text{type}}_{\tau(x_t)} + E^{\text{cls}}_{\kappa(x_t)},$$

where $e(x_t)$is the hierarchical code embedding, $E^{\text{type}}_{\tau(x_t)}$represents the event type, and $E^{\text{cls}}_{\kappa(x_t)}$represents the clinical class.

The hierarchical code embedding combines a token-specific representation with a broader clinical group representation:

$$e(v) = s_v E^{\text{code}}_v + (1 - s_v) E^{\text{group}}_{\text{group}(v)}, s_v = \sigma(\gamma_v)$$

where $s_v$is a learned gate that controls the contribution of the token-specific and group representations. For diagnosis tokens, $\text{group}(v)$ corresponds to the CCSR body system, except for IBD codes, which form a separate group. Non-diagnosis tokens are grouped according to their event type. These assignments produce 30 groups across the token vocabulary. The model used 10 event types, including CCSR diagnosis, IBD codes, laboratory or vital measurements, medications, procedures, sex, age band, race, ethnicity, and special tokens for padding and unknown values. The clinical-class embedding assigned each token to one of eight classes, consisting of disease, symptom,

administrative, injury, neoplasm, pregnancy, IBD, and non-diagnosis. Non-diagnosis events were used as contextual information but were not prediction targets.

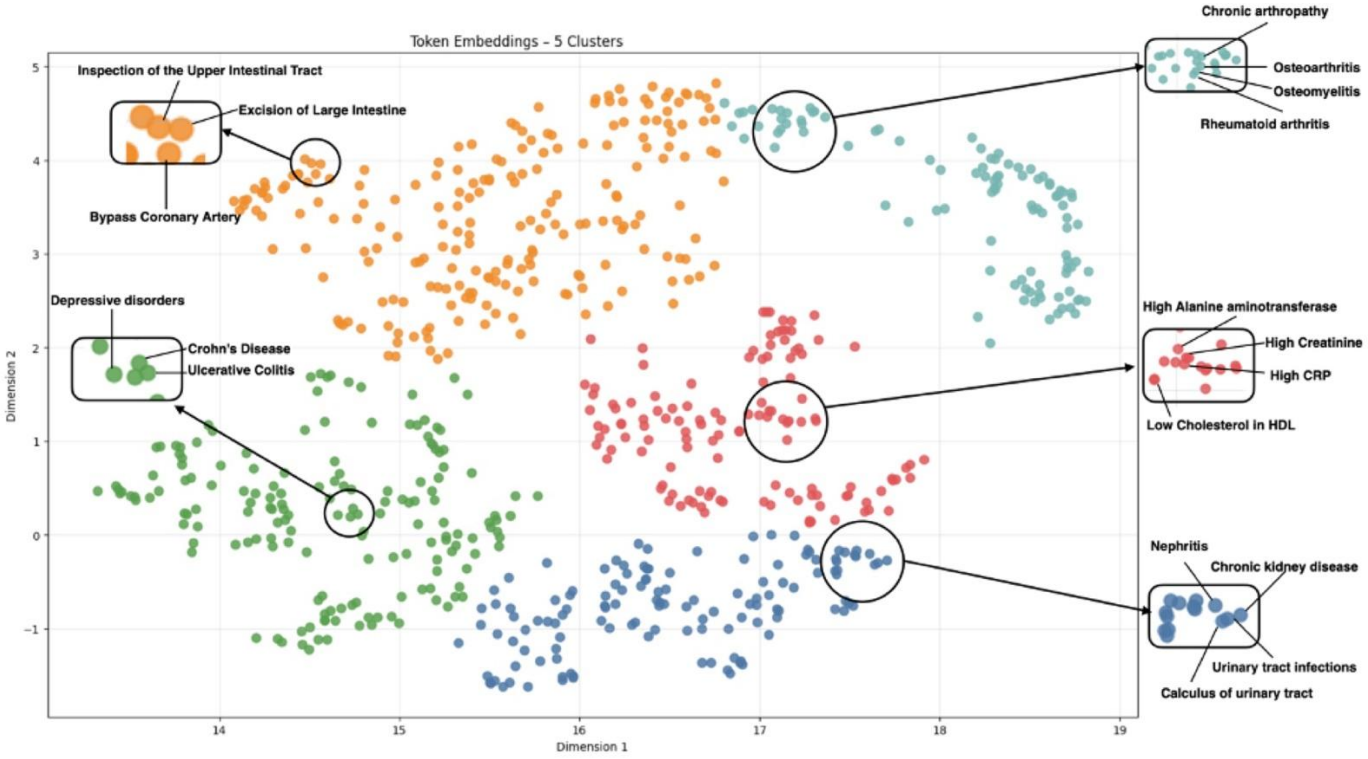


Figure 3. Learned token embeddings showing five distinct clusters of related clinical codes, with similar medical conditions and procedures grouped together in the embedding space

Figure 3 demonstrates that our embedding approach captures meaningful clinical relationships, as evidenced by the clustering of related medical codes. For instance, arthritis-related conditions (osteoarthritis, rheumatoid arthritis) cluster together, while Crohn's disease and ulcerative colitis are positioned close to each other within a gastrointestinal cluster, near depressive disorders. Kidney-related conditions (nephritis, chronic kidney disease, urinary tract infections) group in another cluster, laboratory findings (high C-reactive protein (CRP), high creatinine, high alanine aminotransferase) form a separate cluster, and procedures (excision of large intestine, coronary artery bypass) group together in the fifth cluster.

The resulting representation was processed using rotary positional embeddings and causal masking. Learned event-type biases and age-band decay were incorporated into the attention mechanism to account for relationships between event types and elapsed time. A recurrence mechanism was also included to account for diagnoses previously observed in the patient's history.

For the recurrence mechanism, let $\mathcal{D}_t$denote the diagnosis tokens occurring strictly before position t. Each previous diagnosis received a score:

$$u_{tj} = \frac{(WQht)^T(WKhj)}{\sqrt{d}} + \pi_{x_j}, j \in \mathcal{D}_t,$$

where $\pi_{x_j}$is a learned chronicity prior. The scores were converted to attention weights

$$\alpha_{tj} = \frac{\exp(u_{tj})}{\sum_{j' \in \mathcal{D}_t} \exp(u_{tj'})},$$

which produced a probability distribution over diagnosis codes already observed in the patient's history:

$$P_{\text{copy}}(c \mid x_{\le t}) = \sum_{j \in \mathcal{D}_t} \alpha_{tj} \, \mathbf{1}[m(x_j) = c].$$

The recurrence distribution was combined with the standard vocabulary distribution using a learned gate:

$$g_t = \sigma(w_g^\top h_t + b_g) \cdot \mathbf{1}[\mathcal{D}_t \neq \emptyset].$$

The final probability for diagnosis code $c$was then:

$$z_c = (W_o h_t)_c,$$
$$P(c \mid x_{\le t}) = (1 - g_t)\frac{\exp(z_c)}{\sum_{c'=1}^{586}\exp(z_{c'})} + g_t P_{\text{copy}}(c \mid x_{\le t}).$$

Thus, the model combines its probability distribution across all 586 diagnosis codes with a learned distribution over diagnoses previously observed in the patient. When no earlier diagnosis is present, it uses only the distribution across all diagnosis codes. The input vocabulary contained 725 distinct tokens. At each prediction point, the model assigned a final probability to each diagnosis code and ranked the codes accordingly. The highest-ranked code constituted the Top-1 prediction, while the three, five, and ten highest-ranked codes formed the Top-3, Top-5, and Top-10 predictions, respectively. We also evaluated predictions at the diagnosis-family level, grouping CCSR diagnosis categories by their three-character body-system prefix and treating IBD codes as a separate family. For family match among ranked codes, a prediction was correct if any of the Top-K diagnosis codes belonged to the same family as the observed next diagnosis. For aggregated diagnosis-family ranking, probabilities were first summed across all codes within each family, and families were then ranked by these totals. A prediction was correct if the observed diagnosis family appeared among the Top-K ranked families. Finally, to assess whether predictions captured diagnoses recorded shortly afterward rather than only the immediate next one, we applied a horizon criterion in which a prediction was counted as correct if any of the next H recorded diagnoses (H = 3, 5, or 10) appeared among the Top-K predicted codes.

The training loss was patient-balanced so that patients with different numbers of prediction points contributed equally to model optimization. Patient histories were limited to 256 events, including four demographic tokens and the 252 most recent clinical events. Every disease position served as a prediction point. The model was trained using AdamW with a peak learning rate of 6×10^{-4}, weight decay of 0.1, 5% linear warm-up followed by cosine decay, gradient clipping at 1.0, and a batch size of 128. Dropout of 0.2 and input-token dropout of 15% were used during training. Patients were divided at the patient level into training, validation, and test sets using an 80/10/10 split, corresponding to 90,858, 11,489, and 11,382 patients. Training stopped at epoch 42, and the weights from epoch 22, which achieved the best validation performance, were restored.

2.3 Clinical Interpretation Framework

The agentic framework was developed to extend the predictive transformer from a model that forecasts subsequent diagnoses into a system that can investigate clinically relevant aspects of the predicted trajectory. The framework identifies clinically relevant biomarkers, comorbidities, treatments, and relationships within the available EHR data, determines how these concepts relate to IBD, and directs analytical tools to examine the identified relationships. This separation allows clinical reasoning to guide the analytical questions while ensuring that quantitative results are generated reproducibly by predefined computational procedures.

DeepSeek, accessed through its API as deepseek-chat, selected clinical concepts, proposed relationships, and interpreted aggregate results through task-specific prompts. A predefined workflow linked these tasks to computational procedures for patient selection, sequence modification, model evaluation, and statistical analysis. The implementation provided a LangChain structured-chat executor that exposed the analytical tools through argument schemas and retained intermediate tool calls and results. All analyses used the held-out cohort of 11,382 patients and the frozen transformer. The language model received clinical terminology and aggregate results without patient-level information. Temperature was set to 0.0 for concept selection and relationship assessment and 0.2 for narrative generation.

Biomarker Agent

The Biomarker Agent used the language model's clinical knowledge to identify candidate IBD biomarkers, mapped them to available laboratory codes, and selected eligible measurements for analysis. For each biomarker, it identified patients with a recorded measurement followed by an IBD diagnosis and truncated their histories before that diagnosis. The agent calculated each patient's IBD probability, summed across IBD codes, using the original history and repeated the prediction after masking all occurrences of the selected biomarker. The primary analysis included patients whose baseline Top-1 prediction was an IBD code. Mean and median probabilities before and after masking, together with paired probability differences, summarized the biomarker's contribution, and paired rank-biserial effect sizes characterized the consistency of these differences. Shapley attribution values quantified the biomarker's contribution to individual predictions. The agent compared these contributions between patients whose IBD diagnosis was the Top-1 prediction and those for whom it was not, assessing whether the biomarker carried greater weight in successful predictions. In a separate analysis, features were ranked by absolute Shapley contribution among patients with baseline Top-1 IBD predictions. These analyses characterized how strongly the model's IBD predictions relied on each biomarker, without implying a causal effect on disease development.

Comorbidity Agent

The Comorbidity Agent used the language model's clinical knowledge to identify candidate IBD-associated conditions, mapped them to available diagnosis categories, and selected eligible outcomes for analysis. For each condition, it selected patients with an IBD diagnosis followed by the condition of interest and constructed histories ending immediately before that outcome. The agent compared the probability of the selected condition before and after masking all prior IBD diagnoses. Cohort summaries described the direction, magnitude, and consistency of probability changes, together with changes in diagnosis rank and Top-K inclusion. Shapley values measured the contribution of IBD information and were compared between correctly identified and missed outcomes. These analyses assessed whether prior IBD information supported the model's predictions of subsequent conditions and how that contribution varied across patients.

Counterfactual Agent

The Counterfactual Agent examined how introducing IBD medications into patient histories changed predictions of subsequent comorbidities identified by the Comorbidity Agent. Its language-model component identified candidate medications relevant to IBD management and selected medication–outcome pairs for analysis after checking their representation in the available vocabulary. For each medication–condition pair, the agent identified patients who developed the condition after IBD and had no recorded exposure to the medication in the history used for prediction. It inserted each medication separately after the first IBD diagnosis and compared predictions from the original and modified histories, using only information recorded before the target condition.

This analysis enabled comparison of medication scenarios at both cohort and individual levels. Mean probability changes, confidence intervals, effect sizes, and Top-K changes summarized whether medication insertion increased or decreased predicted comorbidity probabilities across eligible patients. Individual comparisons showed how these responses varied between patients and medications, allowing examination of scenarios obscured by cohort averages. Lower predicted probabilities identified medication–condition relationships for further investigation, including whether a medication might prevent the comorbidity or benefit the individual patient.

Knowledge Extraction Agent

The Knowledge Extraction Agent examined cohort-wide patterns to identify relationships involving IBD, other diagnoses, laboratory measurements, and medications. The language-model component proposed candidate relationships, while analytical tools quantified concept frequency, temporal ordering, and associations. The analysis considered laboratory abnormalities preceding IBD, conditions recorded afterward, and medication use in relation to subsequent diagnoses. Patient counts, effect sizes, and statistical tests were used to identify relationships warranting further investigation.

The agent also investigated connections involving intermediate clinical factors. The language model proposed relationships linking IBD to a subsequent diagnosis through another condition, laboratory abnormality, or medication. Proposals were checked against the available terminology and prediction targets before analysis. Event-ordering tests and comparisons between patient groups assessed the proposed links. This approach examined whether a relationship was supported across several analyses without interpreting an observed sequence as evidence of a causal pathway.

Findings were summarized according to statistical significance after adjustment for multiple comparisons, effect magnitude, and consistency across analyses. The language model interpreted these results and classified relationships as established, emerging, or not established. The resulting report distinguished cohort observations from their proposed clinical interpretation and highlighted candidate relationships for further investigation. Relationships not recognized by the language model were treated as potentially underexplored findings rather than confirmed novel discoveries.

Across agents, probability changes, ranking changes, and attribution characterized model behavior. Analyses evaluated targets at the next 1st, 3rd, 5th, and 10th diagnosis positions, with Top-K comparisons assessing whether the target appeared among the 1, 3, 5, or 10 highest-ranked predictions at each horizon. Paired probability differences were assessed using Wilcoxon signed-rank tests, rank-biserial effect sizes, and bootstrap confidence intervals. After-to-before ratios compared the number of IBD patients in whom a diagnosis was recorded only after versus only before the first IBD record, with p values from an exact binomial (McNemar) test and log-scale 95% confidence intervals. Indirect relationships were assessed by logistic regression adjusted for age band, sex, and log-transformed number of prior events (Wald 95% confidence intervals). Multiple comparisons were adjusted within predefined analytical families using the Benjamini–Hochberg procedure. Shapley contributions were estimated at the next-diagnosis horizon by restoring masked input tokens in randomly sampled orders. Attribution included the available history, including demographic tokens.

3. Results

3.1 Transformer Model Performance

The transformer was evaluated on 318,145 diagnosis prediction positions from 11,382 held-out patients. Exact-code Top-1, Top-3, Top-5, and Top-10 accuracy was 15.28%, 28.69%, 36.67%, and 49.04%, respectively (Table 1). Family matching, which accepted ranked codes from the same diagnosis family as the observed diagnosis, yielded corresponding accuracies of 34.77%, 52.73%, 62.19%, and

74.41%. Aggregating code probabilities within each family before ranking yielded 37.04%, 64.79%, 77.79%, and 92.00%, respectively. These complementary evaluations assessed prediction of the exact next diagnosis and its broader clinical category.

In the broader horizon evaluation, a prediction counted as correct when any of the next H recorded diagnoses appeared among the Top-K predictions. Top-5 accuracy increased from 36.67% for the next diagnosis to 53.09%, 61.36%, and 72.28% across the next three, five, and ten diagnoses, respectively. Corresponding Top-10 values were 49.04%, 68.75%, 77.46%, and 86.52%. These findings suggest that the model captured information about subsequently recorded conditions, even when their exact order was not predicted (Table 1). This pattern may partly reflect noise and variation in EHR documentation, although the broader evaluation also increases the number of diagnoses accepted as correct.

Table 1. Transformer accuracy across prediction ranks

Evaluation endpoint	Top-1 (%)	Top-3 (%)	Top-5 (%)	Top-10 (%)
Exact next diagnosis	15.28	28.69	36.67	49.04
Family match among ranked codes	34.77	52.73	62.19	74.41
Aggregated diagnosis-family ranking	**37.04**	**64.79**	**77.79**	**92.00**
Any diagnosis within next 3	22.80	42.13	53.09	68.75
Any diagnosis within next 5	26.65	49.15	61.36	77.46
Any diagnosis within next 10	33.17	59.58	72.28	86.52

Together, these results suggest that the transformer captured information about subsequent diagnoses and broader clinical patterns in patients' recorded trajectories. Its performance across multiple future diagnoses indicates that the predictions remained informative beyond the immediately next recorded event.

3.2 Automated Discovery of IBD Biomarkers

Among patients whose next recorded diagnosis was IBD and whose highest-ranked prediction was an IBD code, masking inflammatory biomarkers (conventional C-reactive protein, high-sensitivity CRP, and fecal calprotectin) reduced the probability assigned to IBD [25–29]. Median reductions were 1.18 percentage points for conventional CRP, 1.43 for high-sensitivity CRP, and 1.11 for fecal calprotectin. Corresponding mean reductions were 3.02, 3.48, and 1.37 percentage points. High-sensitivity CRP showed the largest reduction by both measures, with mean predicted IBD probability decreasing from 51.67% to 48.19% after masking.
Within the baseline Top-1 subgroup, paired rank-biserial effect sizes were 0.562 for conventional CRP, 0.674 for high-sensitivity CRP, and 0.370 for fecal calprotectin. In its separate full-cohort assessment, the Biomarker Agent classified high-sensitivity CRP and conventional CRP as having a large effect and calprotectin as having moderate effects (Figure 4). The agent characterized these findings as evidence of the model's reliance on biomarker information when predicting IBD, rather than evidence of causal effects on disease.

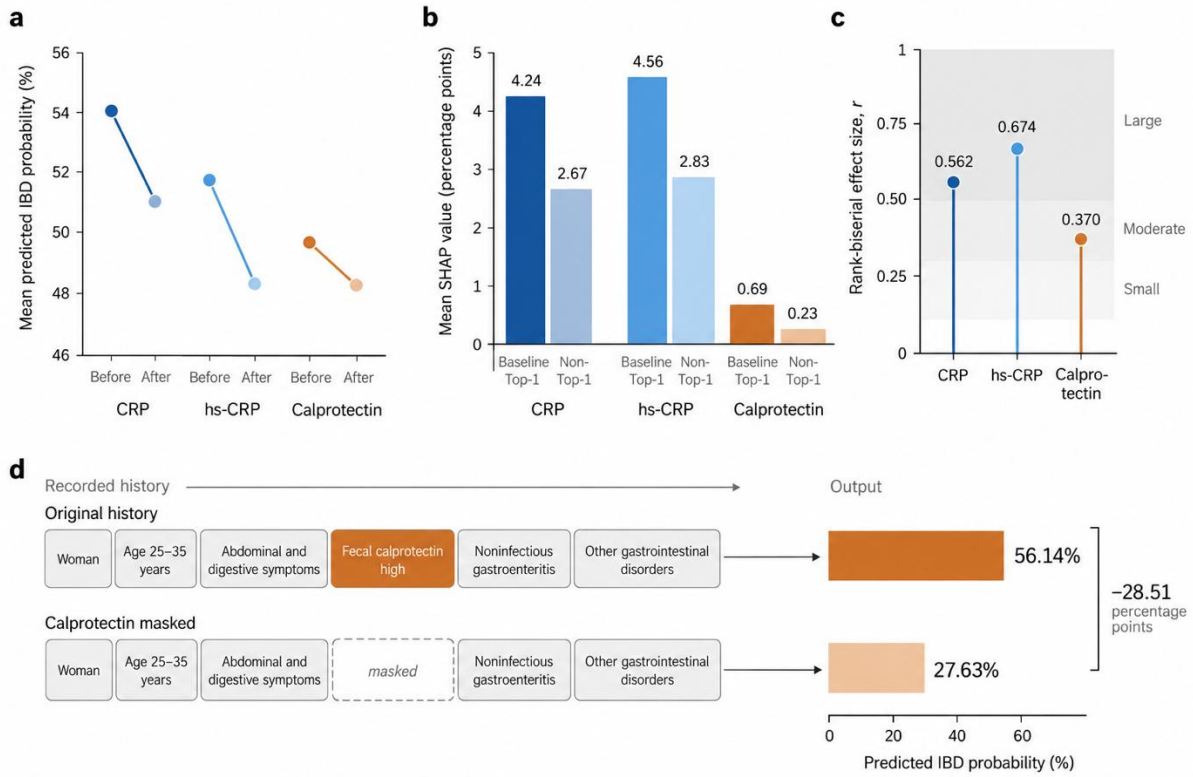


Figure 4. Effects of inflammatory biomarker masking on predicted IBD probability and biomarker contributions. Analyses included held-out patients whose next recorded diagnosis was IBD and whose history contained the corresponding biomarker measurement. Each biomarker was masked separately. **(a)** Mean predicted IBD probability, summed across IBD codes, before and after masking all occurrences of conventional C-reactive protein (CRP), high-sensitivity CRP (hs-CRP), or fecal calprotectin among patients whose baseline Top-1 prediction was an IBD code. Mean reductions were 3.02, 3.48, and 1.37 percentage points, respectively. **(b)** Mean history-based SHAP contributions for baseline Top-1 and sampled non-Top-1 cases: 4.24 versus 2.67 percentage points for CRP, 4.56 versus 2.83 for hs-CRP, and 0.69 versus 0.23 for fecal calprotectin. Baseline Top-1 groups included 714, 152, and 324 patients, respectively; each sampled non-Top-1 group included 200 patients. **(c)** Paired rank-biserial effect sizes for probability reductions in the baseline Top-1 groups (r = 0.562, 0.674, and 0.370, respectively). Shaded bands indicate conventional small, moderate, and large effect-size ranges. **(d)** Individual example of a woman aged 25–35 years whose next recorded diagnosis was Crohn's disease. Her preceding history included abdominal and digestive symptoms, elevated fecal calprotectin, noninfectious gastroenteritis, and other gastrointestinal disorders. With the original history, the frozen model ranked an IBD code first, with a combined probability of 56.14% across IBD codes. Masking all fecal calprotectin occurrences reduced this probability to 27.63%, a decrease of 28.51 percentage points.

Reductions were calculated within patients as baseline minus masked probability, summed across IBD codes, with each biomarker evaluated separately. Mean history-based SHAP contributions were higher among baseline Top-1 cases than among sampled non-Top-1 cases: 4.24 versus 2.67 percentage points for conventional CRP, 4.56 versus 2.83 for high-sensitivity CRP, and 0.69 versus 0.23 for fecal calprotectin. These estimates describe each biomarker's contribution across combinations of recorded information, whereas masking measures its removal from the complete history. Together, the analyses quantified biomarker contributions to IBD predictions from these two perspectives.

Individual-level analysis also illustrated the biomarker contributions observed at the cohort level. A woman aged 25–35 years had a recorded history of abdominal and digestive symptoms, elevated fecal calprotectin, noninfectious gastroenteritis, and other gastrointestinal disorders before a subsequent Crohn's disease diagnosis. With this history, IBD was the model's first prediction, with a combined probability of 56.14% (Figure 4).

Masking calprotectin reduced this probability to 27.63%, a decrease of 28.51 percentage points. Its history-based SHAP contribution was +10.41 percentage points, ranking first among the patient's ten input tokens by absolute attribution magnitude. This case illustrates how patient-level analysis can identify a biomarker contribution exceeding the cohort average while distinguishing attribution across combinations of records from the probability change following masking.

3.3 Automated Discovery of IBD Comorbidities

The Comorbidity Agent identified contributions of IBD history to predictions of clinically recognized complications and associated conditions: intestinal obstruction, particularly in Crohn's disease [20], nutritional deficiencies [22], and IBD-associated inflammatory arthropathies [30]. These were evaluated through the corresponding diagnosis categories of intestinal obstruction and ileus, nutritional deficiencies, and immune-mediated/reactive arthropathies. Among patients whose highest-ranked prediction matched the respective

outcome, masking IBD history reduced predicted probability by a mean of 4.43, 3.67, and 4.98 percentage points, respectively. The corresponding decreases in cohort median predicted probabilities were 2.58, 3.91, and 10.64 percentage points, with paired rank-biserial effect sizes of 0.926, 0.673, and 0.571, respectively (Figure 5).

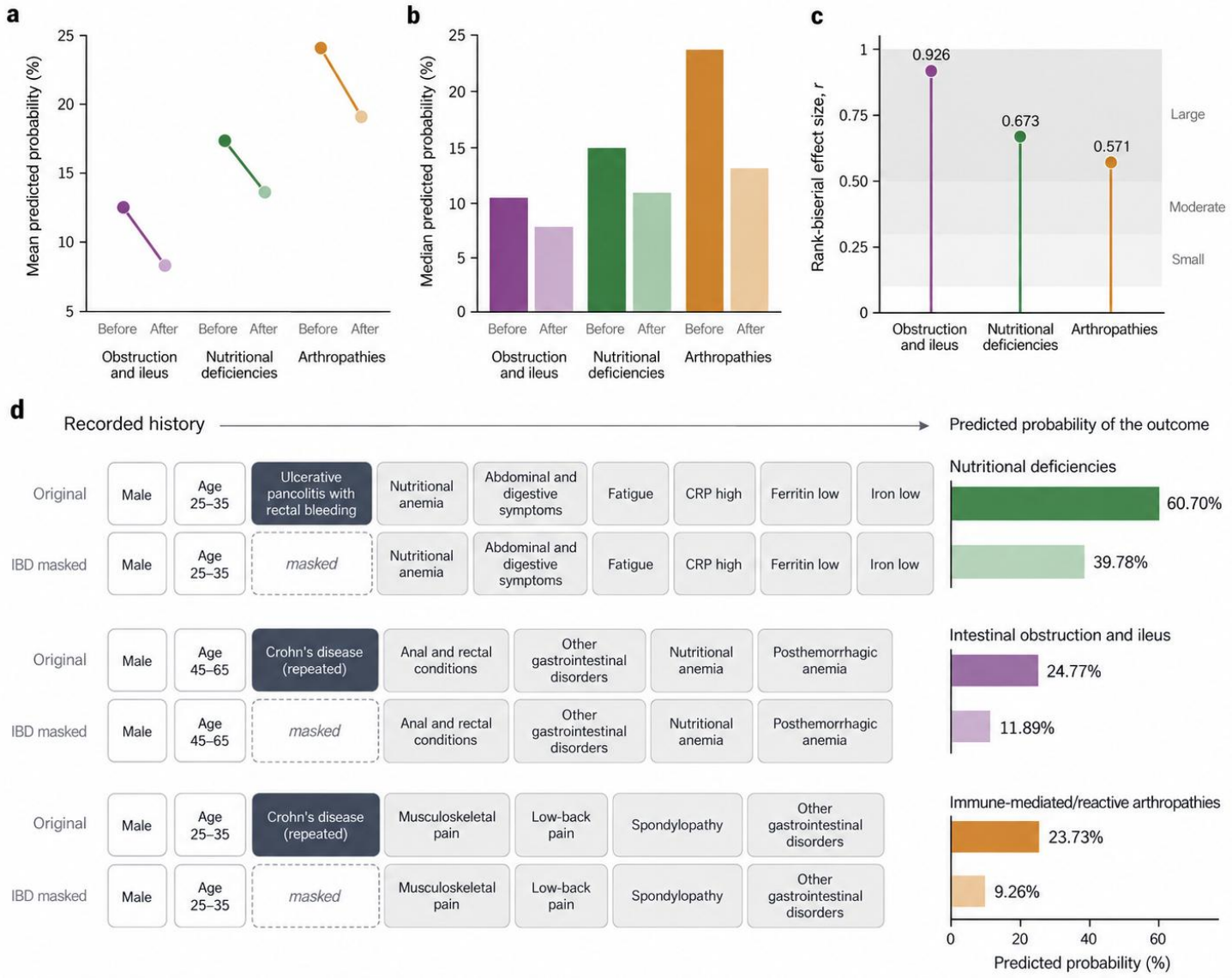


Figure 5. Effects of masking IBD history on predicted probabilities of three IBD-associated comorbidities. Analyses included held-out patients with an IBD diagnosis preceding intestinal obstruction and ileus, nutritional deficiencies, or immune-mediated/reactive arthropathies, whose baseline Top-1 prediction matched the respective outcome. Histories ended immediately before the outcome, and each outcome was evaluated separately. **(a)** Mean predicted probability before and after masking all prior IBD diagnoses. Mean reductions were 4.43, 3.67, and 4.98 percentage points, respectively. **(b)** Median predicted probability before and after masking. **(c)** Paired rank-biserial effect sizes for the probability reductions ($r = 0.926$, 0.673, and 0.571, respectively). Shaded bands indicate small, moderate, and large effect-size ranges. **(d)** Three individual examples comparing original histories with otherwise identical histories in which all IBD diagnoses were masked. Both histories were scored by the frozen transformer; bars show the predicted probability of the indicated outcome. **Top:** A man aged 25–35 years with ulcerative pancolitis with rectal bleeding, nutritional anemia, abdominal and digestive symptoms, fatigue, elevated CRP, low ferritin, and low iron. Masking IBD history reduced the predicted probability of nutritional deficiencies from 60.70% to 39.78%, a decrease of 20.92 percentage points. **Middle:** A man aged 45–65 years with repeated Crohn's disease records, anal and rectal conditions, other gastrointestinal disorders, nutritional anemia, and posthemorrhagic anemia. Masking reduced the predicted probability of intestinal obstruction and ileus from 24.77% to 11.89%, a decrease of 12.88 percentage points. **Bottom:** A man aged 25–35 years with repeated Crohn's disease records, musculoskeletal pain, low-back pain, spondylopathy, and other gastrointestinal disorders. Masking reduced the predicted probability of immune-mediated/reactive arthropathies from 23.73% to 9.26%, a decrease of 14.47 percentage points.

Reductions were calculated within patients as baseline minus masked probability, with each outcome evaluated separately. Mean history-based SHAP contributions from IBD records were higher among baseline Top-1 cases than among sampled non-Top-1 cases: 1.97 versus 0.39 percentage points for intestinal obstruction and ileus, 2.43 versus 0.96 for nutritional deficiencies, and 0.37 versus 0.03 for arthropathies. These positive group means indicate that IBD records contributed to predictions across combinations of available information.

Individual-level analysis examined how IBD history contributed to comorbidity predictions alongside preceding diagnoses and laboratory findings. A man aged 25–35 years had ulcerative pancolitis with rectal bleeding, nutritional anemia, abdominal and digestive symptoms, and fatigue, followed by elevated CRP, low ferritin, and low iron. Nutritional deficiency was the model's first prediction, with a probability of 60.70%. Masking the IBD record reduced this probability to 39.78%, a decrease of 20.92 percentage points (Figure 5). The IBD record had a history-based SHAP contribution of +15.56 percentage points, ranking first among the patient's 16 input tokens by absolute attribution magnitude.

For intestinal obstruction and ileus, a man aged 45–65 years had repeated Crohn's disease records, anal and rectal conditions, other gastrointestinal disorders, and nutritional and posthemorrhagic anemia. The model assigned intestinal obstruction and ileus a probability of 24.77%, ranking it first. Masking the IBD records reduced this probability to 11.89%, a decrease of 12.88 percentage points, and moved the outcome to fourth place. The summed SHAP contribution across the 12 IBD records was +2.29 percentage points, indicating a positive contribution.

For immune-mediated/reactive arthropathies, a man aged 25–35 years had repeated Crohn's disease records, musculoskeletal and low-back pain, a spondylopathy diagnosis, and other gastrointestinal disorders. Arthropathy was the model's first prediction, with a probability of 23.73%. Masking the IBD records reduced this probability to 9.26%, a decrease of 14.47 percentage points, and moved the outcome to second place.

3.4 Counterfactual analysis of medication scenarios

The Counterfactual Agent examined whether adding medication information changed predictions of nutritional deficiency as one of the main comorbidities of IBD. The selected scenarios involved prednisone, used to induce remission in active IBD [31,32]; vedolizumab, an established treatment for ulcerative colitis and Crohn's disease [32,33]; and metronidazole, an antibiotic used in selected IBD-related complications [34,35]. Each medication was inserted separately into histories without prior recorded exposure. Analyses included patients whose baseline first prediction was nutritional deficiency, evaluated one or three diagnosis positions before its recorded occurrence.

At the next-diagnosis position, prednisone insertion reduced mean predicted probability from 18.18% to 16.28%, a decrease of 1.89 percentage points. Metronidazole reduced probability from 17.41% to 15.53%, also a mean decrease of 1.89 points, while vedolizumab reduced it from 17.65% to 16.43%, a decrease of 1.22 points (Table 2). Corresponding paired rank-biserial effect sizes were 0.660, 0.831, and 0.550. When histories ended three diagnosis positions before nutritional deficiency, prednisone reduced mean predicted probability from 14.70% to 13.57%, a decrease of 1.13 percentage points. Metronidazole reduced probability from 14.52% to 12.36%, a decrease of 2.16 points, while vedolizumab reduced it from 14.58% to 13.44%, a decrease of 1.15 points. Corresponding effect sizes were 0.561, 0.929, and 0.599 (Table 2).

Table 2. Medication-insertion effects on nutritional-deficiency predictions at one and three diagnosis positions before the outcome

Medication	Diagnosis positions before outcome	Mean predicted probability of nutritional deficiency, original history (%)	Mean predicted probability of nutritional deficiency, after medication insertion (%)	Rank-biserial effect size
Prednisone	1	18.18	16.28	0.660
Prednisone	3	14.70	13.57	0.561
Metronidazole	1	17.41	15.53	0.831
Metronidazole	3	14.52	12.36	0.929
Vedolizumab	1	17.65	16.43	0.550
Vedolizumab	3	14.58	13.44	0.599

Individual-level analysis enabled comparison of alternative scenarios within the same clinical history. A woman whose latest recorded age band was 45–65 years had hypertension, hemorrhoids, and gastrointestinal hemorrhage documented before ulcerative proctitis. Subsequent records included abdominal and digestive symptoms, unspecified abnormal findings, and further hypertension, hemorrhoid, and gastrointestinal-hemorrhage diagnoses. Nutritional deficiency was the model's first prediction, with a probability of 35.10%. Separate insertion of prednisone, metronidazole, and vedolizumab reduced this probability to 18.34%, 19.25%, and 17.70%, respectively, and moved the outcome from first to second place (Figure 6).

However, individual predictions could change in the opposite direction. A woman aged 35–45 years had an unspecified injury, skin disorders, and skin/subcutaneous symptoms recorded before Crohn's disease involving the small and large intestine, followed by another Crohn's disease record. Nutritional deficiency initially ranked second, with a probability of 11.94%. Separate insertion of prednisone, metronidazole, and vedolizumab increased this probability to 24.51%, 16.53%, and 23.97%, corresponding to increases of 12.57, 4.59,

and 12.03 percentage points. Each insertion moved the outcome to first place (Figure 6). These increases contrasted with the cohort-average reductions, illustrating how individual analysis can identify patient histories with different model responses to the same medication scenarios.

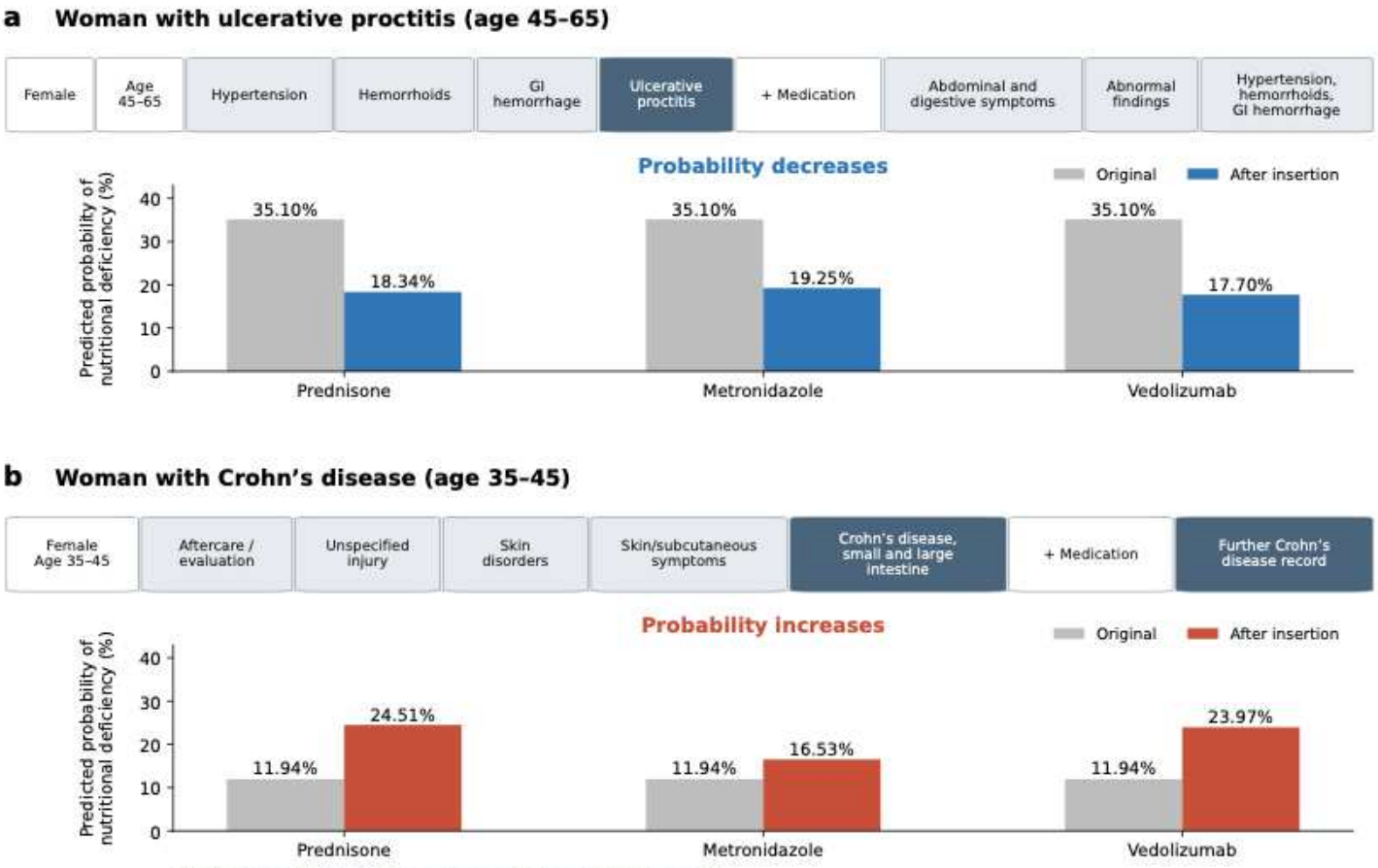


Figure 6. The same medication scenarios change nutritional-deficiency predictions in opposite directions for two patients. Prednisone, metronidazole, and vedolizumab were each inserted separately after the first recorded IBD diagnosis ("+ Medication"). Grey bars show the original predicted probability, and colored bars show the probability after insertion. (a) A woman aged 45–65 years with ulcerative proctitis: insertion reduced probability from 35.10% to 18.34%, 19.25%, and 17.70%, respectively, moving the outcome from first to second place. (b) A woman aged 35–45 years with injury and skin-related records preceding Crohn's disease: insertion increased probability from 11.94% to 24.51%, 16.53%, and 23.97%, respectively, moving the outcome from second to first place in each scenario. Timelines summarize selected records available before prediction; patient (a)'s prior nutritional-deficiency record was retained in the model input but omitted from the displayed timeline

Together, these analyses show how the framework combines group-level summaries with comparisons of medication scenarios for individual patients. The agent interpreted the changes as effects on model predictions. The framework identifies medication scenarios and patient histories for clinical review and further validation, including differences that would be obscured by reporting only average effects.

3.5 Discovery of Novel Patterns

The framework examined direct associations between IBD and subsequent conditions and indirect relationships involving an intermediate diagnosis or laboratory finding. Direct screening identified postoperative digestive complications and opioid-related disorders, both examined previously in IBD populations [36,37]. Digestive complications were more frequently recorded exclusively after than exclusively before the first IBD record, with an after-to-before ratio of 4.35 (95% CI, 3.50–5.41; $p=8.60\times10^{-51}$; $q<0.001$). Opioid-related disorders showed a corresponding ratio of 2.93 (2.10–4.10; $p=2.40\times10^{-11}$; $q<0.001$). These results describe temporal patterns within recorded histories, extending the analysis beyond the outcomes evaluated in the preceding sections.

Indirect analysis identified an association between recorded iron-related laboratory abnormalities and subsequent nutritional anemia among patients with IBD, consistent with the recognized clinical contribution of iron deficiency to anemia in IBD [30]. At the five-diagnosis horizon, low iron was associated with nutritional anemia after adjustment for age band, sex, and prior event count (OR, 19.32; 95% CI, 14.32–26.06; $p<0.001$; $q<0.001$). The association remained evident at the ten-diagnosis horizon (OR, 16.61; 12.46–22.15; $p<0.001$; $q<0.001$). The Knowledge Extraction Agent interpreted the consistent association across horizons, together with the model's response to perturbing iron records, as evidence that this intermediate finding contributed to anemia predictions. A second indirect

relationship connected IBD, intestinal obstruction, and subsequent peritonitis or intra-abdominal abscess [38]. At the five-diagnosis horizon, recorded obstruction was associated with the outcome with an adjusted OR of 7.54 (95% CI, 4.23–13.45; $p=7.42\times10^{-12}$; $q<0.001$). The association persisted at the ten-diagnosis horizon (OR, 6.31; 3.81–10.45; $p=7.64\times10^{-13}$; $q<0.001$). The Knowledge Extraction Agent classified this relationship as supported by both the recorded sequences and model perturbations. This analysis extended the earlier assessment of obstruction as a predicted outcome by examining its role as an intermediate feature associated with a subsequent complication.

Exploratory screening also identified the broad category of benign neoplasms, with an after-to-before ratio of 3.64 (95% CI, 3.02–4.38; $p=6.08\times10^{-51}$; $q<0.001$). Although previous studies have reported associations between IBD and specific colorectal neoplasms [39], our cohort showed an enrichment of the broader benign-neoplasm category after the first recorded IBD diagnosis. The Knowledge Extraction Agent flagged this category for further investigation because its aggregate association was not established as a specific clinical relationship. Together, these examples show how the framework can identify clinically supported relationships and generate more specific questions from additional temporal associations.

4. Discussion

4.1 Principal Findings

This study evaluated whether a longitudinal predictive model and specialized analytical agents could jointly characterize clinically relevant relationships in IBD. Regarding the first research question, the framework identified contributions from established inflammatory biomarkers and IBD-associated conditions. Masking CRP, high-sensitivity CRP, and fecal calprotectin reduced predicted IBD probability, while masking IBD history reduced predictions of intestinal obstruction, nutritional deficiencies, and inflammatory arthropathies. Positive Shapley contributions provided complementary evidence that the model used these records when prioritizing subsequent diagnoses. These findings demonstrate correspondence between selected model associations and clinical knowledge, supporting the framework's ability to examine the clinical information underlying its predictions.

Regarding the second research question, specialized agents connected clinical concepts with predefined analyses and interpretations. The framework translated questions about medications into comparisons of original and modified patient histories. Medication scenarios illustrated this capability, as the insertion of prednisone, metronidazole, and vedolizumab produced average reductions in nutritional-deficiency predictions while individual histories showed changes that differed in magnitude and direction. In this way, the framework can present physicians with several candidate medications alongside the intended effect of each on disease control, leaving the choice of which option to implement to their clinical judgment. The contribution of the agentic framework was therefore to organize clinically motivated investigations and interpret their quantitative outputs.

Regarding the third research question, cohort-scale screening identified direct temporal associations and relationships involving intermediate clinical findings. These analyses recovered previously studied patterns, including postoperative digestive complications and opioid-related disorders, and highlighted broader categories requiring further investigation. The framework linked statistical evidence with model perturbations and agent interpretations, providing a structured approach to prioritizing associations for review. This supports systematic hypothesis generation from longitudinal records.

4.2 Clinical Relevance

Our PRISM framework provides a potential basis for investigating anticipated conditions before they become the focus of clinical care. A physician could examine a predicted diagnosis, identify which recorded biomarkers or preceding conditions contributed to its probability, and assess whether those findings warrant closer monitoring or further evaluation. For example, attribution and masking can show whether an IBD prediction depends on inflammatory measurements or whether a subsequent comorbidity prediction depends on prior IBD information. Presenting these findings within the patient's history may make predictions more interpretable and relevant to longitudinal management.

Medication-scenario analysis adds a complementary way to investigate possible management options. The framework allows alternative medication records to be inserted into the same history and compares the resulting predictions for an individual patient and across eligible patient groups. Cohort summaries describe the overall model response, while individual comparisons reveal differences that averages can conceal. This combination could help physicians formulate questions about treatment selection and prevention in patients with anticipated complications.

The knowledge-extraction component may also support broader clinical investigation by identifying direct associations and connections involving intermediate conditions. Such patterns can help organize questions about monitoring, disease progression, and possible intervention points. Computational scenario screening could prioritize medication–outcome hypotheses before resource-intensive clinical studies, allowing many candidate scenarios to be explored without exposing patients to experimental treatments.

4.3 Prior Work

Earlier longitudinal models established that clinical sequences contain information useful for predicting subsequent events. Doctor AI used recurrent networks to forecast clinical events, while BEHRT and Med-BERT developed transformer-based representations for disease prediction [5–7]. Foresight extended this approach to generative modeling of patient timelines [9]. These developments broadened the range of outcomes that could be anticipated from patient histories. Their application to disease management also requires understanding which clinical findings support a prediction and how that information might guide monitoring.

Interpretability research has examined how explanations can help clinicians recognize potentially preventable events. Lundberg and colleagues developed a system that predicted intraoperative hypoxaemia and identified contributing patient and procedural factors [40]. Trained on records from more than 50,000 surgeries, the system improved anaesthesiologists' prediction performance in an experimental evaluation. The authors estimated that assistance could increase the proportion of anticipated hypoxaemia events from 15% to 30%. This study connected prediction explanations with potentially modifiable factors, illustrating how interpretation can support consideration of preventive action.

Clinical implementation studies have further examined whether predictive information can support timely treatment. In a prospective multisite evaluation of the TREWS sepsis warning system, clinician confirmation of an alert within three hours was associated with an adjusted absolute reduction in in-hospital mortality of 3.3 percentage points (95% CI, 1.7–5.1), corresponding to an 18.7% relative reduction [41]. These observational findings highlight the importance of connecting predictions to clinical assessment and treatment workflows. The evaluated systems addressed defined acute events; a complementary opportunity is to connect anticipated diagnoses with their contributing history and alternative management scenarios across chronic disease trajectories.

PRISM builds on these approaches by integrating clinically guided attribution, intervention comparisons, and direct and indirect association screening around a longitudinal predictive model. Combining cohort summaries with individual analyses enables examination of both shared patterns and patient-specific changes in predicted outcomes within one workflow.

4.4 Limitations and Future Works

Several limitations define the scope of this study. IBD provided a clinically diverse use case, but application to other diseases and healthcare settings remains to be evaluated. Medication scenarios were constrained by the available treatment information, including limited detail on dose, duration, adherence, and concurrent therapies. The available modalities also provided an incomplete representation of each patient's clinical state, limiting the physiological and treatment-related information available for individual scenario analysis.

Future work can extend the framework to other diseases and independent datasets and incorporate procedures and combined interventions. Integrating genetics, imaging, clinical notes, longitudinal measurements, and detailed treatment histories could support a more comprehensive patient representation, moving toward a partial digital twin for exploring predicted responses to different interventions. Comparing these simulated responses with observed outcomes would establish their relevance to treatment planning. Prospective physician evaluations should examine how cohort summaries and individual analyses support interpretation, monitoring, and management decisions, alongside their effects on workflow, patient outcomes, and resource use.

5. Conclusion

This study combined a longitudinal predictive transformer with specialized agents to investigate biomarkers, comorbidities, medication scenarios, and clinical associations in IBD. The model achieved an exact-code Top-10 accuracy of 49.04%, increasing to 86.52% when predictions were evaluated against any of the next ten recorded diagnoses. Masking and attribution analyses identified contributions from established inflammatory biomarkers and demonstrated how IBD history supported predictions of IBD comorbidities. These findings show that the framework could connect anticipated diagnoses with clinically relevant information in patient histories and quantify its contribution at both cohort and individual levels.

Inserting prednisone, metronidazole, or vedolizumab reduced mean predicted nutritional-deficiency probability by 1.89, 1.89, and 1.22 percentage points, respectively, with reductions of up to 17.40 points. Knowledge extraction identified postoperative digestive complications and opioid-related disorders more frequently after than before IBD, with ratios of 4.35 and 2.93, respectively (both $q<0.001$), alongside an indirect relationship linking IBD, intestinal obstruction, and subsequent peritonitis or intra-abdominal abscess.

6. Declaration

6.1 CRediT authorship contribution statement

Mohammad Assadi Shalmani: Conceptualization, Methodology, Data curation, Software, Validation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Masoud Khani:** Conceptualization, Methodology, Data curation,

Software, Validation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Preetika Sinh:** Data curation, Validation, Writing – review & editing. **Qiang Lu:** Validation, Writing – review & editing. **Jake Luo:** Conceptualization, Data curation, Validation, Supervision, Writing – review & editing.

6.2 Funding

This research received no external funding.

6.3 Data Availability Statement

The de-identified patient data used in this study were accessed through the TriNetX research network. Restrictions apply to the availability of these data, which were used under license for the current study and are therefore not publicly available. Information on accessing the platform is available on the TriNetX website. Summary data and statistical analysis results are available from the corresponding author upon reasonable request.

6.4 Ethical Considerations

This study was approved by the Institutional Review Board of the Medical College of Wisconsin and Froedtert Hospital (IRB# PRO00045896). The data analyzed consisted of fully de-identified patient records; therefore, informed consent was not required in accordance with institutional and federal regulations.

6.5 Conflicts of Interest

The authors declare that there are no financial or non-financial interests that could directly or indirectly influence or be perceived to influence the work submitted for publication. Each author has thoroughly examined the manuscript's content and confirmed that any affiliations, financial involvement, or personal connections do not present a conflict of interest in the design, execution, interpretation, or reporting of the research findings.

6.6 AI-assisted Technologies in the Manuscript Preparation Process

During the preparation of this work, the authors used Anthropic Claude and OpenAI ChatGPT for language editing and grammatical check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final manuscript.

7. References

[1] Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017;24(1):198–208. doi: 10.1093/jamia/ocw042.

[2] Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform. 2018;22(5):1589–1604. doi: 10.1109/JBHI.2017.2767063.

[3] Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1:18. doi: 10.1038/s41746-018-0029-1.

[4] Singh A, Nadkarni G, Gottesman O, Ellis SB, Bottinger EP, Guttag JV. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. J Biomed Inform. 2015;53:220–228. doi: 10.1016/j.jbi.2014.11.005.

[5] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. Proc Mach Learn Res. 2016;56:301–318. PMID:28286600.

[6] Li Y, Rao S, Solares JRA, et al. BEHRT: transformer for electronic health records. Sci Rep. 2020;10:7155. doi: 10.1038/s41598-020-62922-y.

[7] Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021;4:86. doi: 10.1038/s41746-021-00455-y.

[8] Amirahmadi A, Ohlsson M, Etminani K. Deep learning prediction models based on EHR trajectories: a systematic review. J Biomed Inform. 2023;144:104430. doi: 10.1016/j.jbi.2023.104430.

[9] Kraljevic Z, Bean D, Shek A, et al. Foresight, a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. Lancet Digit Health. 2024;6(4):e281–e290. doi: 10.1016/S2589-7500(24)00025-6.

[10] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195. doi: 10.1186/s12916-019-1426-2.

[11] Jung K, Kashyap S, Avati A, et al. A framework for making predictive models useful in practice. J Am Med Inform Assoc. 2021;28(6):1149–1158. doi: 10.1093/jamia/ocaa318.

[12] Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67. doi: 10.1038/s42256-019-0138-9.

[13] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–4774. Available from: https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[14] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745–e750. doi: 10.1016/S2589-7500(21)00208-9.

[15] van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. Eur J Epidemiol. 2020;35(7):619–630. doi: 10.1007/s10654-020-00636-1.

[16] de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med. 2022;5:2. doi: 10.1038/s41746-021-00549-7.

[17] Kent DM, Paulus JK, van Klaveren D, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern Med. 2020;172(1):35–45. doi: 10.7326/M18-3667.

[18] Lin L, Sperrin M, Jenkins DA, Martin GP, Peek N. A scoping review of causal methods enabling predictions under hypothetical interventions. Diagn Progn Res. 2021;5:3. doi: 10.1186/s41512-021-00092-9.

[19] Vavricka SR, Schoepfer A, Scharl M, Lakatos PL, Navarini A, Rogler G. Extraintestinal manifestations of inflammatory bowel disease. Inflamm Bowel Dis. 2015;21(8):1982–1992. doi: 10.1097/MIB.0000000000000392.

[20] Malik S, Farooq U, Schwartz DA, Adler DG. Analysis of patients with Crohn's disease and intestinal obstruction: a National Inpatient Sample study. Ann Gastroenterol. 2024;37(5):543–551. doi: 10.20524/aog.2024.0911.

[21] Thia KT, Sandborn WJ, Harmsen WS, Zinsmeister AR, Loftus EV Jr. Risk factors associated with progression to intestinal complications of Crohn's disease in a population-based cohort. Gastroenterology. 2010;139(4):1147–1155. doi: 10.1053/j.gastro.2010.06.070.

[22] Bischoff SC, Bager P, Escher J, et al. ESPEN guideline on clinical nutrition in inflammatory bowel disease. Clin Nutr. 2023;42(3):352–379. doi: 10.1016/j.clnu.2022.12.004.

[23] Karreman MC, Luime JJ, Hazes JMW, Weel AEAM. The prevalence and incidence of axial and peripheral spondyloarthritis in inflammatory bowel disease: a systematic review and meta-analysis. J Crohns Colitis. 2017;11(5):631–642. doi: 10.1093/ecco-jcc/jjw199.

[24] Sinh P, Cross RK. Cardiovascular comorbidities and inflammatory bowel disease: causes and consequences. Gastroenterol Hepatol (N Y). 2024;20(4):204–215. PMID:38682122.

[25] Ananthakrishnan AN, Adler J, Chachu KA, et al. AGA clinical practice guideline on the role of biomarkers for the management of Crohn's disease. Gastroenterology. 2023;165(6):1367–1399. doi: 10.1053/j.gastro.2023.09.029.

[26] Singh S, Ananthakrishnan AN, Nguyen NH, et al. AGA clinical practice guideline on the role of biomarkers for the management of ulcerative colitis. Gastroenterology. 2023;164(3):344–372. doi: 10.1053/j.gastro.2022.12.007.

[27] Menees SB, Powell C, Kurlander J, Goel A, Chey WD. A meta-analysis of the utility of C-reactive protein, erythrocyte sedimentation rate, fecal calprotectin, and fecal lactoferrin to exclude inflammatory bowel disease in adults with IBS. Am J Gastroenterol. 2015;110(3):444–454. doi: 10.1038/ajg.2015.6.

[28] van Rheenen PF, Van de Vijver E, Fidler V. Faecal calprotectin for screening of patients with suspected inflammatory bowel disease: diagnostic meta-analysis. BMJ. 2010;341:c3369. doi: 10.1136/bmj.c3369.

[29] Kiss LS, Papp M, Lovasz BD, et al. High-sensitivity C-reactive protein for identification of disease phenotype, active disease, and clinical relapses in Crohn's disease: a marker for patient classification? Inflamm Bowel Dis. 2012;18(9):1647–1654. doi: 10.1002/ibd.21933.

[30] Gordon H, Burisch J, Ellul P, et al. ECCO guidelines on extraintestinal manifestations in inflammatory bowel disease. J Crohns Colitis. 2024;18(1):1–37. doi: 10.1093/ecco-jcc/jjad108.

[31] Ko CW, Singh S, Feuerstein JD, et al. AGA clinical practice guidelines on the management of mild-to-moderate ulcerative colitis. Gastroenterology. 2019;156(3):748–764. doi: 10.1053/j.gastro.2018.12.009.

[32] Feuerstein JD, Ho EY, Shmidt E, et al. AGA clinical practice guidelines on the medical management of moderate to severe luminal and perianal fistulizing Crohn's disease. Gastroenterology. 2021;160(7):2496–2508. doi: 10.1053/j.gastro.2021.04.022.

[33] Singh S, Loftus EV Jr, Limketkai BN, et al. AGA living clinical practice guideline on pharmacological management of moderate-to-severe ulcerative colitis. Gastroenterology. 2024;167(7):1307–1343. doi: 10.1053/j.gastro.2024.10.001.

[34] Barnes EL, Agrawal M, Syal G, et al. AGA clinical practice guideline on the management of pouchitis and inflammatory pouch disorders. Gastroenterology. 2024;166(1):59–85. doi: 10.1053/j.gastro.2023.10.015.

[35] Nguyen GC, Loftus EV Jr, Hirano I, Falck-Ytter Y, Singh S, Sultan S. American Gastroenterological Association Institute guideline on the management of Crohn's disease after surgical resection. Gastroenterology. 2017;152(1):271–275. doi: 10.1053/j.gastro.2016.10.038.

[36] Cohen BL, Fleshner P, Kane SV, et al. Prospective cohort study to investigate the safety of preoperative tumor necrosis factor inhibitor exposure in patients with inflammatory bowel disease undergoing intra-abdominal surgery. Gastroenterology. 2022;163(1):204–221. doi: 10.1053/j.gastro.2022.03.057.

[37] Cohen-Mekelburg S, Rosenblatt R, Gold S, et al. The impact of opioid epidemic trends on hospitalised inflammatory bowel disease patients. J Crohns Colitis. 2018;12(9):1030–1035. doi: 10.1093/ecco-jcc/jjy062.

[38] Jeong SH, Choi JS, Kim JW, et al. Clinical features of intra-abdominal abscess and intestinal free-wall perforation in Korean patients with Crohn's disease: results from the CONNECT study. J Clin Med. 2021;10(1):116. doi: 10.3390/jcm10010116.

[39] Axelrad JE, Olén O, Söderling J, et al. Inflammatory bowel disease and risk of colorectal polyps: a nationwide population-based cohort study from Sweden. J Crohns Colitis. 2023;17(9):1395–1409. doi: 10.1093/ecco-jcc/jjad056.

[40] Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2(10):749–760. doi: 10.1038/s41551-018-0304-0.

[41] Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med. 2022;28(7):1455–1460. doi: 10.1038/s41591-022-01894-0.