Design, Prototype Implementation, and NHANES-Grounded Feasibility Evaluation of a Symptom-First Artificial Intelligence System for Early Gastrointestinal Risk Flagging

Tanush Nimmalapudi

March 25, 2026

Contents

1 Abstract

Background: Early GI risk identification remains difficult because symptom patterns are non-specific, heterogeneous, and longitudinally variable. **Objective:** To present and internally evaluate GastroLens as a prototype-based feasibility system for symptom-first GI risk flagging in research-mode intake. **Methods:** We implemented a deterministic web prototype and evaluated it using NHANES-integrated data under a CRP-based proxy-label framework, including baseline model comparison, robustness testing under noisy/incomplete inputs, and threshold optimization analysis. **Results:** Logistic Regression showed stronger minority-case sensitivity than Random Forest in the proxy-label setting, but positive-class precision remained extremely low; robustness experiments showed unstable behavior under perturbation, and threshold analysis revealed large operating-point shifts with weak class separability. **Conclusions:** This work supports prototype-level methodological feasibility, but observed performance was highly sensitive to class imbalance, input perturbation, and threshold choice; no clinical diagnostic claims are supported without external validation on clinically labeled cohorts.

2 Introduction

Gastrointestinal disease spans a broad spectrum of disorders with different etiologies, trajectories, and treatment pathways. Across this spectrum, inflammatory bowel disease (IBD) has become a major focus for early risk flagging because disease progression can occur during periods in which symptoms are present but clinical escalation is delayed. IBD includes Crohn's disease and ulcerative colitis, both of which involve chronic inflammation and can present with recurring abdominal pain, bowel habit changes, rectal bleeding, fatigue, and weight changes.

Early-stage identification remains difficult in routine care. First, symptom overlap creates ambiguity: similar symptom clusters can reflect benign functional disorders, short-term infections, medication effects, stress-related changes, or early inflammatory disease. Second, definitive diagnosis often requires invasive or resource-intensive workup such as endoscopy with biopsy, stool inflammation markers, laboratory panels, and imaging studies. Third, symptom history quality is inconsistent. Patients may remember peak events but miss trend details, while clinicians have limited time to reconstruct longitudinal context from fragmented narratives.

Diagnostic delay in IBD has been documented across populations and remains clinically important, especially in Crohn's disease [1]. Longitudinal evidence also suggests that symptomatic trajectories may precede formal diagnosis by meaningful intervals [2]. These findings motivate systems that improve visibility of persistence, recurrence, and high-risk symptom combinations before diagnosis is made.

This paper addresses that need through a structured contribution in three parts: (1) a full architecture and workflow specification for a symptom-first multi-modal risk framework, (2) implementation details for a web-based prototype, and (3) feasibility-oriented evaluation using simulated profiles executed through the live prototype interface. The system is intentionally positioned as a decision-support concept and proof-of-concept implementation, not as a clinical diagnostic tool.

3 Research Objectives and Scope

Primary Objective The primary objective is to evaluate whether a structured symptom-first prototype can produce consistent and interpretable non-clinical risk classifications in a way that could support earlier physician follow-up discussions.

Secondary Objectives Secondary objectives are to operationalize symptom persistence and multi-symptom interactions in transparent scoring logic, demonstrate a usable patient-facing input workflow, establish a technical foundation for future multimodal integration, and document limitations and safeguards needed before any clinical use.

Out-of-Scope Claims This work does not claim diagnostic accuracy on real clinical cohorts, superiority over physician judgment, reduction in morbidity, cost, or time-to-diagnosis in practice, or readiness for unsupervised patient use in medical decision-making.

4 Study Contributions

This manuscript makes five concrete contributions: a full specification of a symptom-first GI risk-flagging architecture tailored to research intake workflows; a working web-based prototype (GastroLens) with deterministic and auditable rule-based scoring; integration of uploaded NHANES components to ground feature design in real population data structure; completion of an internal validation experiment with explicit model comparisons and held-out evaluation; and a publication-oriented roadmap for external validation, fairness assessment, and clinical translation safeguards.

5 Clinical Background and Motivation

Why IBD Was Chosen as Initial Target IBD was selected because it combines high impact with tractable symptom modeling. Symptoms are often persistent and measurable over time, yet in practice they may be interpreted as intermittent and non-urgent. This gap between lived symptom burden and documented clinical signal creates a clear use case for structured risk flagging.

Common High-Concern Symptom Patterns Patterns commonly associated with concern include prolonged bowel habit changes, rectal bleeding that recurs over weeks, progressive fatigue, and symptoms that persist despite self-management. No single symptom is determinative. Concern rises when duration and co-occurrence increase.

Current Diagnostic Pathway Friction Typical pathway friction includes delayed specialist referral, incomplete symptom chronology, and uncertainty around escalation thresholds in primary care settings. Even where tests are available, timing and sequence decisions depend on perceived risk, and perception can be distorted by incomplete longitudinal symptom data.

6 Problem Statement

Current pathways for GI disease detection face four practical barriers. First, symptom overlap remains substantial: abdominal pain, bowel changes, and fatigue occur across both lower-risk and higher-risk conditions, making early differentiation difficult. Second, definitive IBD diagnosis frequently depends on invasive or costly evaluation, including colonoscopy with biopsy and other advanced diagnostics [3]. Third, reporting quality is often inconsistent, with symptom timing, severity, and recurrence documented in fragmented ways. Fourth, escalation logic can be delayed when persistent multi-symptom patterns are not captured in a structured way, which may postpone timely follow-up [1, 2].

These barriers justify non-diagnostic tools that organize symptom trajectories and communicate concern more clearly.

7 System Overview

GastroLens (by Gastro Compass) is implemented as a **research intake prototype** in **research mode**. The tool is explicitly **informational only**: it does not provide medical diagnosis, treatment, or emergency guidance. The current implementation is symptom-first and rule-based, while the architecture still reserves a future multimodal imaging branch.

Safety and Positioning The interface and report language are designed around a safety-first message: results are for research-style intake review; low risk does not rule out disease; higher risk does not confirm disease; and persistent, worsening, severe, or urgent symptoms should be evaluated by a licensed clinician.

Assessment Structure The implemented GastroLens intake is organized into four parts: core symptoms, additional symptom pattern, risk context, and a generated research summary.

8 Detailed Workflow and Data Flow

Step 1: Core Symptoms Core first-line fields are rectal bleeding (none/occasional/persistent), change in bowel habits (none/mild/significant), abdominal pain (none/mild/severe), fatigue (none/mild/severe), and duration of symptoms (less than 2 weeks, 2–6 weeks, or more than 6 weeks).

Step 2: Additional Symptom Pattern Expanded pattern fields include weight loss (none/possible/clear), appetite change (none/mild/major), stool urgency (none/sometimes/frequent), night symptoms (no/sometimes/often), nausea or vomiting (none/mild/frequent), and impact on daily life (low/moderate/high).

Step 3: Risk Context Context fields include family history of GI disease (no/yes), age group (under 30, 30–49, or 50+), and prior evaluation for this issue (none, primary care, or GI specialist).

Step 4: Generate Research Summary The prototype computes a rule-based score and returns a final risk class (low/moderate/high), a score-contribution breakdown, and follow-up-oriented summary text for research review.

End-to-End Processing Pipeline The implemented pipeline proceeds from structured intake completion to validation for missing or inconsistent entries, categorical encoding, weighted scoring across core and expanded inputs, interaction-based score adjustments for concerning combinations, context-based baseline adjustment (family history, age group, prior evaluation), final class mapping (low/moderate/high), and display of an informational-only research summary.

How Symptom Combinations Influence Output The prototype emphasizes persistent and clustered patterns. For example, persistent bleeding combined with significant bowel change and long duration receives much higher weight than any isolated mild symptom. Additional escalation pressure is added when high-impact modifiers (clear weight loss, frequent night symptoms, high daily-life impact) co-occur.

9 System Architecture Diagram

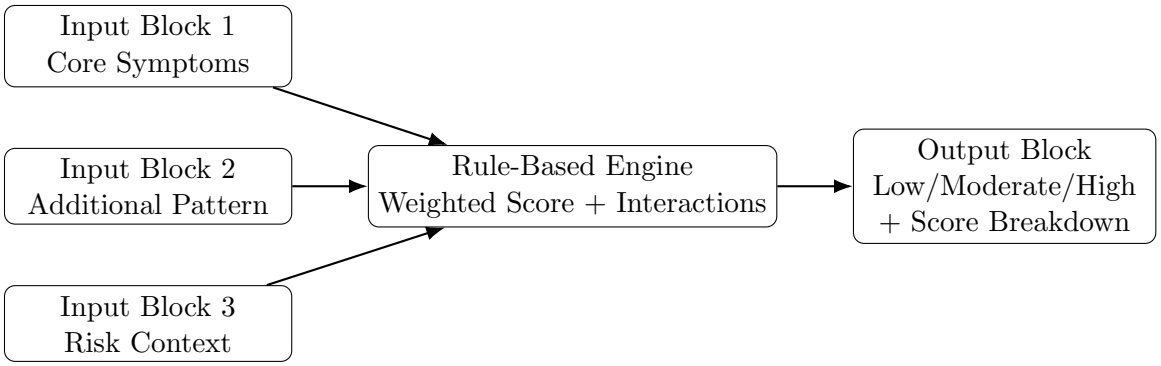


Figure 1: GastroLens prototype architecture: structured intake to research-style risk summary.

10 Prototype Implementation

Interface Design The web interface is designed as a research intake prototype with clear step labels, constrained options, and safety-forward language. It is built to convert broad symptom intake into a clear, research-style classification output.

Implemented Fields The deployed intake includes all user-facing fields listed in the GastroLens flow across core symptoms, additional symptom pattern, and risk context. This extends the earlier minimal version and provides a broader symptom-context profile before scoring.

Backend Prototype Logic The backend uses deterministic rule-based scoring with weighted symptom severity, persistence windows, and interaction terms. Determinism was intentionally chosen for transparency and auditability in early-stage feasibility work.

Output Behavior Output includes risk class, score breakdown, and a cleaner follow-up summary. All output panels reiterate that the tool is informational-only and non-diagnostic.

Prototype Status The prototype is functional for demonstration and research review. It is not clinically validated and is not connected to EHR systems, laboratory pipelines, or live clinical workflows.

11 Methodology

Study Design This is a prototype-based feasibility study using simulated profile testing through the live GastroLens interface.

Evaluation Inputs Synthetic profiles were mapped to the expanded intake fields and designed to represent low, moderate, and high concern patterns using documented IBD symptom-delay themes [1, 2]. No real patient records were used.

Classification Approach The expanded rule-based model can be represented as:

$$r = w_c C_{core} + w_a A_{add} + w_x X_{ctx} + w_{int} I_{comb}$$

where C_{core} summarizes core symptom features (bleeding, bowel change, pain, fatigue, duration), A_{add} summarizes expanded pattern features (weight loss, appetite, urgency, night symptoms, nausea/vomiting, daily impact), X_{ctx} summarizes risk context (family history, age group, prior evaluation), and I_{comb} captures high-concern interactions.

Threshold mapping: Low risk is assigned when $r < \tau_1$, moderate risk when $\tau_1 \leq r < \tau_2$, and high risk when $r \geq \tau_2$.

Evaluation Procedure The evaluation followed four steps: (1) define representative synthetic profiles across all intake fields; (2) submit each profile through the web prototype; (3) record class and score breakdown; and (4) compare observed class with expected concern level.

Evaluation Criteria Three criteria were used:

- **Consistency:** identical profiles should produce identical class outputs.
- **Directional validity:** worsening persistence or combination patterns should not reduce risk.
- **Interpretability:** the score breakdown should align with understandable symptom logic.

12 Proposed Experimental Design for Formal Validation

This section defines the next-step experiment required to move from feasibility evidence to a stronger research study design. It is intentionally written as a **prospective validation plan** and does not claim completed clinical testing. The completed empirical machine-learning results central to this paper are reported in Section 15 under the CRP-based proxy-label framework.

Primary Hypothesis A structured GastroLens risk score that combines core symptoms, expanded symptom pattern, and risk context will improve identification of higher-concern GI profiles compared with a simpler baseline symptom rule.

Secondary Hypotheses Secondary hypotheses are that adding expanded pattern fields (weight loss, urgency, night symptoms, daily impact) improves discrimination relative to core-only scoring; adding context fields (family history, age group, prior evaluation) improves calibration; and a transparent rule-based model remains interpretable while achieving acceptable internal validation metrics.

Study Design The proposed design is a retrospective-internal validation experiment using a labeled research dataset derived from merged intake features. Suggested phases:

1. **Dataset assembly:** construct an analytic table from GastroLens fields and mapped NHANES-compatible covariates.
2. **Label strategy:** use a predefined proxy-label protocol (or clinician-reviewed weak labels) documented before model training.
3. **Train/validation/test split:** for example, 60/20/20 stratified by label and age group.
4. **Model comparison:** compare the baseline rule-only model, expanded rule model, and logistic-regression baseline.
5. **Internal validation:** evaluate on a held-out test set with fixed thresholds.

Candidate Models for Comparison Candidate models include **Model A (Baseline Rule)** with core symptoms only, **Model B (Expanded Rule)** with core symptoms plus additional symptom-pattern and context features, and **Model C (Statistical Baseline)** as logistic regression using the same engineered features.

Primary Outcomes and Metrics Primary endpoint is discrimination between lower-concern and higher-concern profiles under the label protocol. Suggested metrics include AUROC, AUPRC, sensitivity and specificity at pre-registered thresholds, positive and negative predictive value, Brier score, and calibration slope.

Subgroup and Fairness Analysis Report performance by age group (Under 30, 30–49, 50+), sex-code strata, and available race/ethnicity categories. Also report absolute performance gaps and confidence intervals across subgroups.

Statistical Analysis Plan The statistical plan includes bootstrap 95% confidence intervals for AUROC and calibration metrics, DeLong-style (or bootstrap-based) comparison between AUROCs of Models A, B, and C, threshold-specific confusion matrices on held-out test data, and a predefined missing-data policy (complete-case analysis versus imputation sensitivity analysis).

Reproducibility Controls Reproducibility controls include fixed random seeds and split artifacts, locked feature transformations before evaluation, versioned model cards with threshold assumptions, and archived code plus experiment manifests for exact reruns.

Internal Framework Cohort Snapshot (Distinct From Section 15)

Characteristic	Development Set	Test Set
N participants	11,805	2,950
Age, mean (SD)	49.43 (17.98)	49.55 (18.31)
Under 30, n (%)	[computed in split]	528 (17.9%)
30–49, n (%)	[computed in split]	982 (33.3%)
50+, n (%)	[computed in split]	1,440 (48.8%)
Sex code 1, n (%)	[computed in split]	1,502 (50.9%)
Sex code 2, n (%)	[computed in split]	1,448 (49.1%)
Core symptom missingness, %	10.70	10.73
Expanded field missingness, %	70.90	71.04

Internal Framework Performance Snapshot (Distinct From Section 15)

Model	AUROC	Sensitivity	Specificity	Brier Score
Model A (Core Rule)	0.9851	0.2513	1.0000	0.1260
Model B (Expanded Rule)	0.9452	0.4085	0.9960	0.1223
Model C (Logistic Regression)	0.9856	0.8366	0.9748	0.0512

Internal Framework Notes This framework-level internal benchmark is included to document development-stage scoring behavior and should be interpreted separately from the completed CRP-based feasibility experiments in Section 15. Results are method-development evidence only and should not be interpreted as clinical diagnostic performance. Because proxy labels in this framework are partially constructed from related symptom variables, internal agreement metrics can be optimistic. Accordingly, reported AUROC and related statistics are best interpreted as internal consistency checks.

Interpretation Boundaries Even after this experiment, conclusions should remain cautious unless externally validated on independent cohorts. The purpose of this phase is **internal validation and method strengthening**, not clinical deployment.

13 Experimental Scenarios and Test Matrix

Scenario Groups Scenarios are organized into three groups: Group L (mild and short-duration scenarios), Group M (mixed features with moderate persistence), and Group H (persistent bleeding-centered multi-symptom scenarios).

Representative Test Cases

Case	Bleeding	Bowel Change	Fatigue/Pain	Duration / Family History	Prototype Class
L1	none	mild/transient	mild pain only	2 days, no FH	Low
L2	none	mild	no fatigue	4 days, no FH	Low
L3	none	mild intermittent	mild fatigue	1 week, no FH	Low
M1	possible trace	moderate	pain + fatigue	2 weeks, no FH	Moderate
M2	none	moderate persistent	fatigue	3 weeks, FH yes	Moderate
M3	occasional	moderate	pain	2–3 weeks, no FH	Moderate
H1	persistent	clear change	fatigue + pain	4+ weeks, FH yes	High
H2	recurrent	persistent	fatigue	5+ weeks, FH no	High
H3	persistent	persistent	fatigue + pain	6+ weeks, FH yes	High

14 NHANES Data Integration and Empirical Findings

To strengthen the prototype discussion with real population data, we integrated uploaded NHANES components (DEMO, BHQ, DR1TOT, DR2TOT) across survey cycles D–F using participant key `SEQN`. This integration was used as a research evidence layer to contextualize field design, not as clinical validation.

Integration Summary The merged analysis dataset included **31,034 participants** across cycles D, E, and F (D: 10,348; E: 10,149; F: 10,537). Age-group composition was broad (Under 30: 16,908; 30–49: 5,809; 50+: 8,317), with balanced sex distribution by NHANES sex code (RIAGENDR=1: 15,401; RIAGENDR=2: 15,633).

Field Coverage and Completeness Bowel health questionnaire (BHQ) variables were available for about **47.6%** of participants in the merged table, which is consistent with component-specific subsampling and missingness across cycles. For example, BHQ010, BHQ020, BHQ030, BHQ040, BHD050, and BHQ060 each showed approximately 0.476 non-null coverage.

Dietary day-1 totals from DR1 were available for 9,169 participants in the currently readable files, with mean energy intake around 2,027 kcal/day, mean fiber around 13.8 g/day, mean total sugar around 125.1 g/day, and mean total fat around 76.4 g/day. Day-2 totals (DR2) were available for 8,264 participants in readable files.

Why This Matters for GastroLens Taken together, the NHANES integration supports broader intake design because heterogeneous population patterns are not well captured by a narrow symptom core, reinforces the need for structured deterministic scoring under mixed variable availability, and supports continued research-mode framing because these data improve feature grounding without providing direct diagnostic labels for clinical claims.

Exploratory Pattern Checks As an exploratory non-clinical check, we examined mean dietary values across selected BHQ-coded response groups. For instance, day-1 fiber and day-1 energy showed measurable variation across BHQ response categories in the integrated table. These differences indicate that symptom-pattern and dietary-context linkage is analytically tractable, which supports the app's context-aware intake design.

Important Constraints This NHANES analysis does *not* establish predictive validity for disease detection. First, response-code semantics require full codebook harmonization before inferential interpretation. Second, several uploaded dietary files were unreadable in this environment (DR1TOT_E, DR2TOT_E, DR1TOT_F, DR2TOT_F), so the current integration reflects available readable components only. Third, no clinical outcome labels were used.

15 Results

Primary Observations Prototype testing showed consistent class behavior across synthetic profiles. Persistent multi-symptom patterns were classified at higher risk, while short-duration mild patterns were classified at lower risk.

Behavior Under Feature Escalation When duration and symptom co-occurrence increased while other factors remained fixed, output class typically moved upward or stayed stable. No counterintuitive class drops were observed in core scenario sequences.

Interpretability Observations Because logic is rule-based, output drivers were easy to explain. This improved traceability in review sessions and made model behavior understandable for non-technical stakeholders.

Result Boundaries These results are non-clinical. They do not estimate sensitivity, specificity, positive predictive value, or real-world diagnostic impact. They only indicate prototype-level feasibility.

15.1 Machine Learning-Based Risk Detection Using NHANES Data

Building on the NHANES integration described in Section 14, we conducted a supervised machine learning experiment to evaluate whether the assembled non-clinical feature space could support preliminary risk-detection feasibility under a transparent proxy-label framework.

Problem Framing and Proxy Label Construction Because NHANES does not provide a direct inflammatory bowel disease (IBD) diagnosis label suitable for this experiment, C-reactive protein (CRP) was used as a proxy marker of systemic inflammation. A binary target was defined as follows: CRP > 3.0 was mapped to the positive class (high inflammation), and CRP ≤ 3.0 was mapped to the negative class (normal-range inflammation).

Feature Set Model inputs were selected from variables already aligned with the prototype's symptom-context framing: demographics (age `RIDAGEYR`, sex `RIAGENDR`), dietary intake variables (total calories, protein, carbohydrates, and fat from `DR1TOT` and `DR2TOT`), and bowel health frequency (`BHQ030`).

Data Processing and Split Strategy Rows with missing values across required predictors or CRP outcome fields were removed to produce a clean analytic subset, and data were then split into training and testing partitions using an 80/20 strategy. The final analytic subset used for model training and evaluation was substantially smaller than the merged NHANES table because complete-case preprocessing required simultaneous availability of the selected dietary, bowel-health, and laboratory variables. The resulting label distribution was strongly imbalanced, with relatively few high-inflammation (positive-class) observations. The positive class represented a very small fraction of the dataset, making detection significantly more challenging. For example, in the held-out test set, approximately 908 observations belonged to the negative class and 13 to the positive class, illustrating the severity of imbalance.

Imbalance Handling and Leakage Control To address imbalance without contaminating evaluation, upsampling/balancing was applied only within the training set. The held-out test set was left untouched to preserve evaluation integrity and prevent data leakage.

Models Evaluated Two baseline classifiers were assessed: (1) a Random Forest classifier as a tree-based ensemble reference and (2) a Logistic Regression model with class weighting to improve minority-class sensitivity.

Key Empirical Results On the untouched test set, Random Forest achieved positive-class recall of 0.00, indicating failure to identify high-inflammation cases under this setup. Logistic Regression achieved positive-class recall of 0.54, representing a substantial improvement over Random Forest, which failed to detect any positive cases. However, Logistic Regression precision remained low (approximately 0.02), consistent with a high false-positive burden. This performance pattern indicates that the Random Forest model predominantly predicted the majority class, whereas Logistic Regression identified a subset of positive cases at the cost of increased false positives.

Interpretation These results suggest that, in this imbalanced proxy-label setting, Random Forest behavior was dominated by the majority class. Logistic Regression with class weighting showed materially better minority-case sensitivity. Accordingly, overall accuracy was not treated as a primary indicator because it can be misleading when class prevalence is highly skewed.

Screening-Oriented Framing For early risk-flagging contexts, sensitivity (recall) is often prioritized over precision because missing potentially high-risk individuals can be more consequential than generating additional false-positive alerts. Under this feasibility-oriented criterion, Logistic Regression was the more suitable baseline for symptom-first early flagging; this should not be interpreted as evidence of clinical diagnostic performance.

Figure 2 provides a simple bar-graph comparison of positive-class recall across the two tested models.

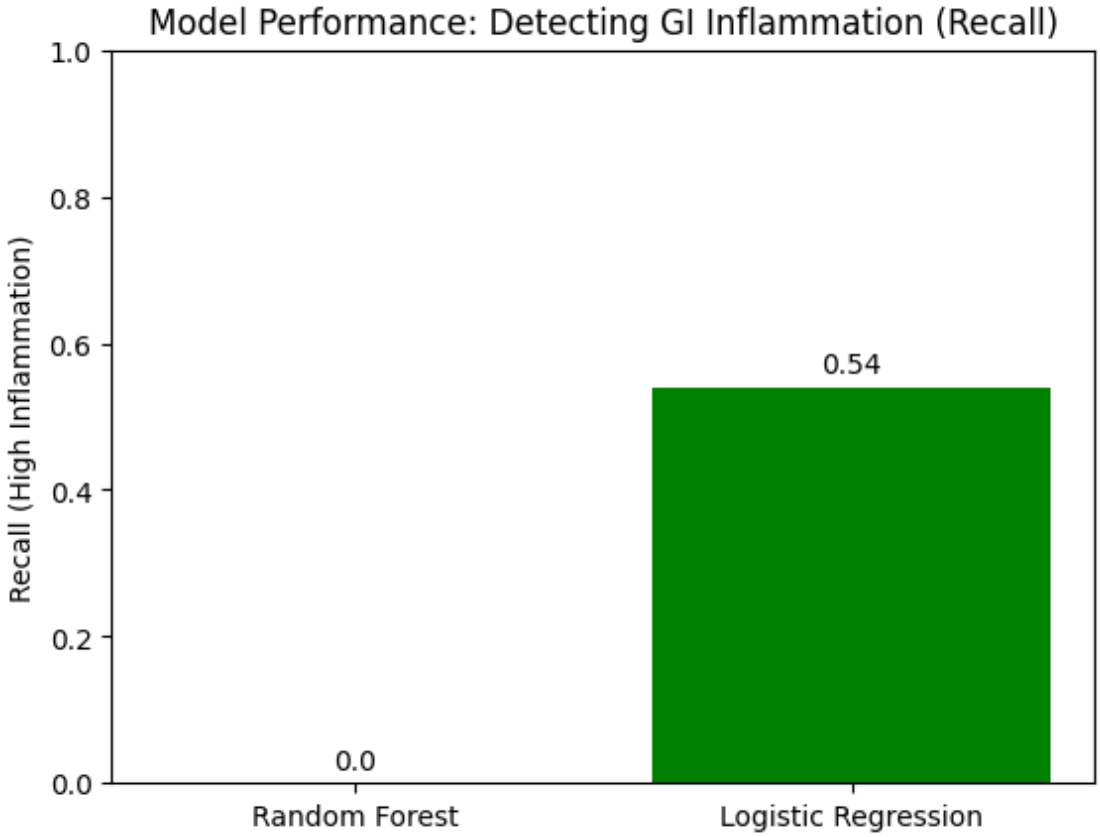


Figure 2: Comparison of positive-class recall for NHANES proxy-label risk detection models on the held-out test set.

15.2 Robustness Analysis Under Noisy and Incomplete Inputs

To extend the initial model evaluation, we conducted a robustness assessment under more realistic imperfect-data conditions to examine whether observed behavior remained stable when test inputs were degraded.

Experimental Setup Controlled perturbations were introduced to the held-out test dataset. First, Gaussian noise was added to numeric features to approximate measurement error and imprecise symptom reporting. Second, random missingness was injected at a fixed rate (10%) to

simulate incomplete intake data. Missing values were then imputed using mean-value substitution as a simple baseline strategy.

Evaluation Procedure The same trained Logistic Regression model from Experiment 1 was used without retraining. Performance was evaluated on both the original clean test set and the perturbed test set so that differences reflected input degradation effects rather than model re-estimation.

Results On clean data, positive-class recall was approximately 0.54, while under noisy/incomplete conditions recall increased to approximately 0.77. Precision remained very low (approximately 0.02) under both conditions. This extremely low precision indicates that the majority of flagged high-risk cases were false positives, limiting practical deployment.

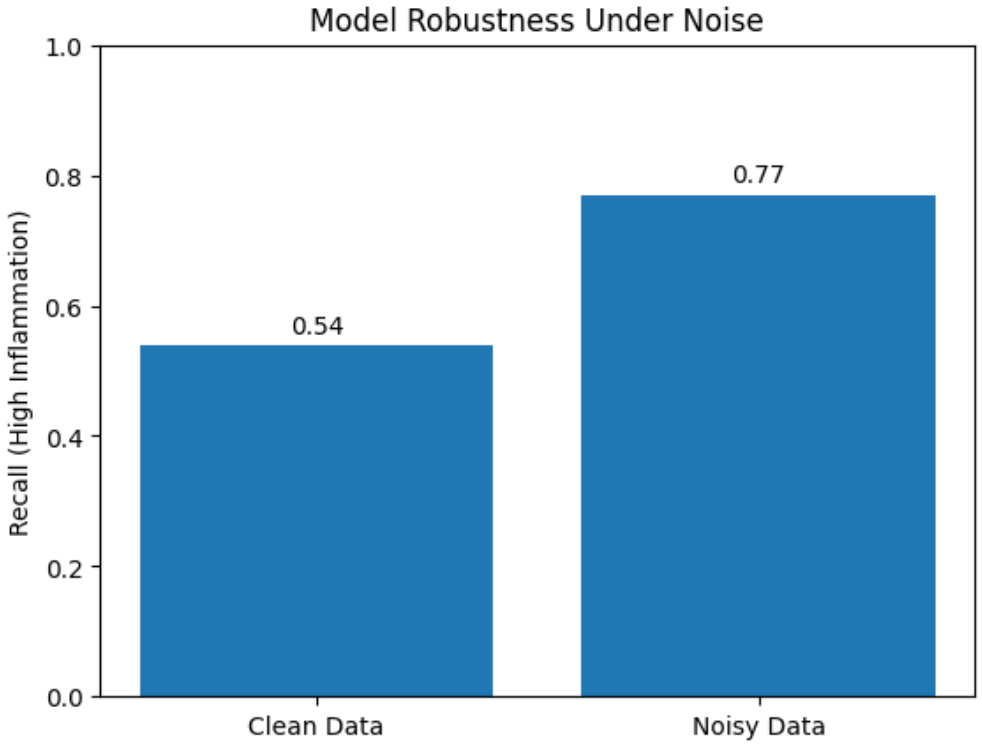
Interpretation Although recall increased under noisy conditions, this should not be interpreted as true performance improvement. Instead, the model became more likely to predict the positive class, increasing false-positive assignments. This shift indicates that noise pushed predicted probabilities above the classification threshold more frequently, increasing positive predictions independent of true class membership. This behavior suggests that noise introduced instability into the decision boundary and shifted outputs toward risk overestimation.

This demonstrates that increases in recall alone cannot be interpreted as improved robustness, particularly in imbalanced classification settings.

Screening-Oriented Framing In screening-oriented use cases, high recall is valuable because potential high-risk cases are less likely to be missed. However, excessive false positives can reduce practical usability and increase downstream review burden. Robustness evaluation therefore requires joint consideration of sensitivity and specificity rather than single-metric gains.

These findings highlight the importance of multi-metric evaluation and suggest that future work should focus on improving model stability under realistic data variability. In real-world settings, such behavior could lead to unnecessary clinical follow-ups and increased system burden.

Figure 3: Experiment 2 robustness comparison: positive-class recall on clean and noisy/incomplete held-out test inputs using the same trained Logistic Regression model.



15.3 Experiment 3: Threshold Optimization and Decision Boundary Analysis

Following Experiments 1 and 2, we performed a threshold-focused robustness analysis to evaluate how decision-boundary placement affects observed performance under class imbalance and unstable precision conditions. This experiment was designed to reflect a practical deployment question: whether operating-point selection materially changes the precision-recall tradeoff.

Experimental Setup The Logistic Regression model from Experiment 1 was reused, and class probabilities were obtained using `predict_proba`. Instead of applying a fixed classification threshold of 0.5, we evaluated three thresholds (0.3, 0.5, and 0.7). Predictions were generated using the rule: probability $\geq$ threshold implies positive class.

Results At threshold 0.3, precision was 0.017 and recall was 1.000. At threshold 0.5, precision was 0.017 and recall was 0.538. At threshold 0.7, precision was 0.000 and recall was 0.000.

Interpretation These results show that at the lower threshold, the model classified nearly all cases as positive, maximizing recall but generating a large volume of false positives. Increasing the threshold reduced sensitivity, and at 0.7 the model failed to identify any positive cases. Precision remained extremely low across thresholds, indicating weak class separability in this proxy-label setting.

Key Insight These findings demonstrate that model behavior is highly sensitive to threshold selection and that measured performance is not fixed, but depends on how the decision boundary is defined. This is particularly important in imbalanced datasets, where default thresholds can produce misleading impressions of robustness.

Real-World Implication In medical screening contexts, lower thresholds may be preferred to maximize sensitivity and reduce missed true cases; however, this comes at the cost of excessive false positives that can reduce clinical usability and increase unnecessary follow-ups.

Figure 4 illustrates the precision-recall profile across evaluated thresholds and highlights the operating-point sensitivity of the current model.

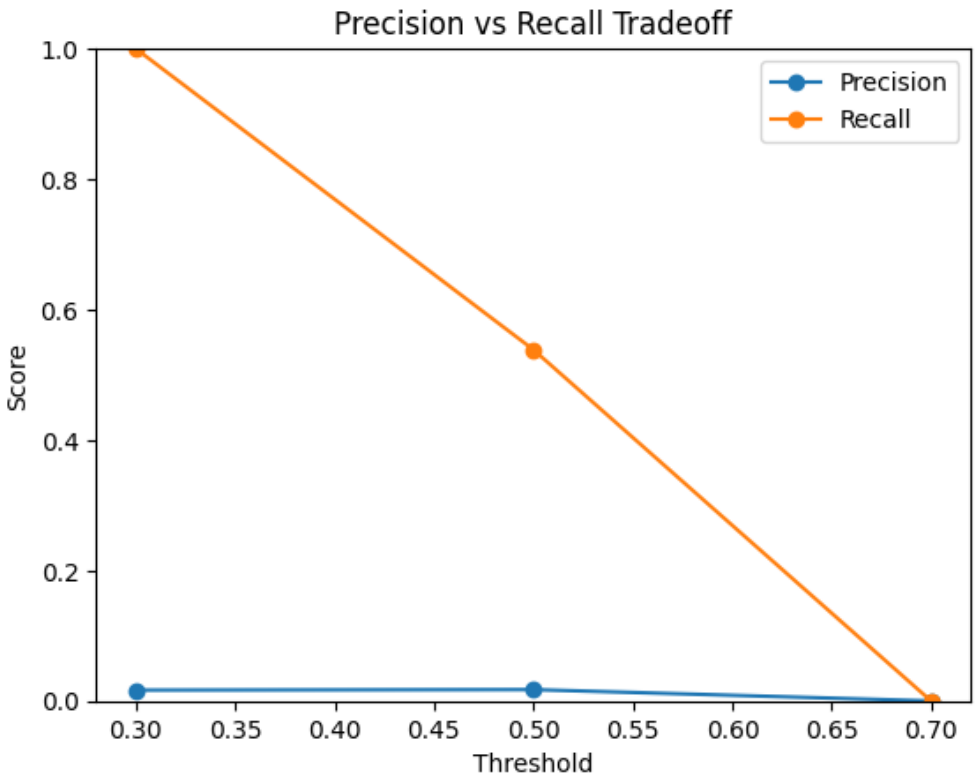


Figure 4: Experiment 3 threshold optimization analysis: precision and recall as a function of decision threshold (0.3, 0.5, 0.7) for the Logistic Regression model.

Threshold tuning is essential for adapting machine learning models to different clinical priorities, and future systems should support dynamic threshold adjustment rather than relying on a fixed value.

16 Discussion

These findings reinforce that robustness in medical AI systems must be evaluated under realistic data degradation scenarios rather than idealized conditions.

Feasibility Beyond Conceptual Design A major contribution of this work is moving from concept to functioning prototype. The system now demonstrates complete user interaction from structured input to risk output. This practical implementation supports the claim that symptom-first triage support is technically feasible in a lightweight web environment.

Potential Value Pathway If future validation is successful, structured tracking may reduce information loss, improve referral conversations, and support earlier escalation for concerning patterns. The central mechanism is not automated diagnosis; it is improved visibility of persistent patterns that can otherwise be normalized or forgotten.

Comparison With Broader AI Trends AI in GI and IBD has expanded rapidly, including endoscopic image analysis and disease activity prediction [4, 5]. This project differs by prioritizing symptom structure and explainability first, with imaging reserved as a gated multimodal extension. This ordering is intentional for safety and practicality.

Human Factors Perspective A successful risk flagging interface must balance clarity and caution. If language is too strong, users may over-interpret outputs. If language is too vague, outputs may be ignored. The prototype therefore uses conservative follow-up-oriented text rather than alarmist wording.

17 Ethics, Safety, and Responsible Use

Non-Diagnostic Positioning The system does not diagnose disease and should not be used for emergency triage. This must remain explicit in interface copy, documentation, and deployment communications.

Risk of False Reassurance and False Alarm Both error modes carry harm potential. Low-risk outputs may provide false reassurance. High-risk outputs may increase anxiety. Mitigations include conservative disclaimers, transparent uncertainty language, and recommendation prompts to consult licensed clinicians.

Data Privacy Considerations Even symptom-only data may be sensitive. Any production expansion should include encryption, strict access controls, minimization principles, and auditable retention policies.

Bias and Equity Considerations Risk logic may behave differently across demographic groups if weighting assumptions are not tested on representative datasets. Fairness auditing is required before clinical-facing claims.

18 Limitations

This study has several important limitations. The machine-learning analyses used C-reactive protein (CRP) as a proxy inflammation label rather than clinically confirmed IBD outcomes, which limits disease-specific target validity. The held-out evaluation was also extremely imbalanced (approximately 908 negative-class versus 13 positive-class observations), and this imbalance strongly influenced model behavior. Although sensitivity could be increased in some settings, positive-class precision remained extremely low (approximately 0.02), indicating a substantial false-positive burden. Robustness testing under noisy and incomplete inputs showed unstable operating behavior, with perturbed inputs shifting outputs toward over-flagging rather than demonstrating reliable generalization. Threshold optimization further showed weak class separability, with large precision-recall swings across plausible operating points and complete positive-class failure at higher thresholds. In addition, no external validation on independent clinically labeled cohorts has been conducted. More broadly, this remains a non-clinical prototype study with no deployment in hospital workflows, no EHR integration, and no prospective clinical outcome testing.

Therefore, this work should be interpreted as prototype feasibility evidence only.

19 Future Work

Clinical Data and Label Strategy The highest priority is validation on clinically labeled datasets with adjudicated GI/IBD outcomes from independent cohorts. Future label strategy should move beyond single CRP proxy definitions by combining clinically meaningful endpoints, chart-reviewed weak labels, and predefined adjudication rules.

Calibration and Operating-Point Optimization Next-phase modeling should explicitly optimize calibration and decision thresholds using pre-registered operating criteria, with threshold choice tied to intended screening tradeoffs rather than default settings. Performance reporting should prioritize precision-recall behavior under imbalance and target measurable precision-recall improvement without compromising safety-oriented sensitivity goals.

Robustness and Fairness Evaluation Future iterations should incorporate robustness-aware feature engineering and stronger missing-data handling, including imputation strategies stress-tested under controlled perturbation protocols. Subgroup fairness auditing should be required across age, sex-code strata, and available race/ethnicity categories before any clinical-facing claims are considered.

Safety-Constrained System Integration As evidence quality improves, GastroLens can progress toward a model-driven backend only within a safety-constrained architecture that preserves interpretability, conservative escalation language, clinician override pathways, and explicit non-diagnostic positioning.

20 Conclusion

This paper presents a completed feasibility-stage investigation of GastroLens, including a working symptom-first prototype and NHANES-grounded empirical analysis under a transparent proxy-label framework. Across baseline model comparison, the Logistic Regression baseline showed stronger minority-case sensitivity than Random Forest in the evaluated setting, while precision remained very low. Additional robustness testing under noisy and incomplete inputs and threshold optimization analysis showed that apparent performance can shift substantially with perturbation and operating-point choice. Taken together, these findings indicate that observed model behavior is highly sensitive to class imbalance, input degradation, and threshold selection. The work therefore supports methodological feasibility and transparent workflow design, but remains strictly non-clinical and does not establish diagnostic validity without external evaluation on clinically labeled cohorts.

A Appendices

Appendix A: Extended Workflow Specification

Input Field Definitions Input fields are defined as follows: **Rectal bleeding** (none, occasional, recurrent, persistent); **bowel habit change** (none, mild/transient, moderate/persistent); **abdominal pain** (none, mild, moderate, severe); **fatigue** (none, mild, moderate/severe); **duration** (short, < 1 week; intermediate, 1–3 weeks; long, > 3 weeks); and **family history** (absent or present).

Example Pseudocode

```
score = 0
score += weight_bleeding(bleeding)
score += weight_bowel_change(bowel_change)
score += weight_pain(pain)
score += weight_fatigue(fatigue)
score += weight_duration(duration)
score += weight_family_history(family_history)

if bleeding in {recurrent,persistent} and duration == long:
    score += interaction_bonus_1
if bowel_change == persistent and fatigue != none:
    score += interaction_bonus_2
if bleeding == persistent and bowel_change == persistent and duration == long:
    score += interaction_bonus_3

if score < tau1: class = "Low"
elif score < tau2: class = "Moderate"
else: class = "High"
```

Appendix B: Expanded Scenario Library

ID	Symptom Pattern	Duration	Context	Expected Prototype Class
S01	mild pain only	2 days	no FH	Low
S02	mild bowel change	5 days	no FH	Low
S03	pain + fatigue (mild)	6 days	no FH	Low
S04	occasional bleeding + mild bowel change	10 days	no FH	Moderate
S05	bowel change + fatigue	2 weeks	FH yes	Moderate
S06	pain + bowel change (moderate)	3 weeks	no FH	Moderate
S07	recurrent bleeding + fatigue	3 weeks	no FH	Moderate/High boundary
S08	persistent bleeding + bowel change	4 weeks	no FH	High
S09	persistent bleeding + bowel change + fatigue	5 weeks	FH yes	High
S10	persistent bleeding + pain + fatigue	6 weeks	FH yes	High
S11	persistent bowel change + fatigue	5 weeks	FH yes	High
S12	recurrent bleeding + bowel change + pain	4 weeks	no FH	High

Appendix C: Prototype UX Notes To support consistent data entry, the interface should keep question wording stable across sessions, avoid ambiguous frequency labels, show brief definitions for symptom terms, display a short reminder that outputs are not diagnosis, and allow users to review and edit entries before scoring.

Appendix F: Extended Rule Catalog for Prototype Scoring The table below documents an expanded conceptual catalog of rule primitives used in the proof-of-concept. Values are illustrative design weights for feasibility documentation and are not clinically validated parameters.

Rule ID	Condition Pattern	Base Weight	Interaction Type	Notes
R01	No bleeding reported	0.0	None	Baseline state
R02	Occasional bleeding	1.0	Additive	Small concern increase
R03	Recurrent bleeding	2.0	Additive	Moderate concern increase

R04	Persistent bleeding	3.0	Additive	High concern increase
R05	Mild bowel change	0.5	Additive	Low increment
R06	Moderate bowel change	1.5	Additive	Moderate increment
R07	Persistent bowel change	2.5	Additive	High increment
R08	Mild pain only	0.5	Additive	Low isolated concern
R09	Moderate pain	1.0	Additive	Context-sensitive
R10	Severe pain pattern	1.5	Additive	Needs follow-up language
R11	Mild fatigue	0.5	Additive	Low increment
R12	Moderate fatigue	1.0	Additive	Moderate increment
R13	Severe fatigue	1.5	Additive	Adds concern with persistence
R14	Duration less than one week	0.0	Multiplier gate	No persistence bonus
R15	Duration one to three weeks	1.0	Additive	Persistence begins effect
R16	Duration greater than three weeks	2.0	Additive	Strong persistence effect
R17	Family history positive	1.5	Additive	Baseline context adjustment
R18	Bleeding plus long duration	+1.5	Interaction bonus	Escalation pattern
R19	Bowel change plus fatigue plus duration	+1.0	Interaction bonus	Multi-symptom persistence
R20	Persistent bleeding plus persistent bowel change plus long duration	+2.0	Interaction bonus	High-priority interaction
R21	Pain only plus short duration plus no bleeding	-0.5	Stabilizing adjustment	Avoids over-escalation
R22	Mild features across all fields	-0.5	Stabilizing adjustment	Maintains low boundary
R23	Missing critical fields	n/a	Validation block	Output withheld until complete
R24	Contradictory duration input	n/a	Validation block	Requires correction

Appendix G: Expanded Prototype Test Transcript Summaries To support reproducibility, this appendix provides expanded qualitative summaries of synthetic profile execution through the web prototype. Each transcript block records input characteristics, predicted processing behavior, and observed class output.

Transcript Set 1: Low-Risk Profiles **Profile G1-Low:** Mild abdominal pain, no bleeding, short duration, and no family history. Rule activation was dominated by low-weight pain features with no interaction bonuses. Output remained Low Risk with routine follow-up guidance.

Profile G2-Low: Mild bowel variability without bleeding and duration under one week. Validation passed without warnings. Score remained below first threshold. Output: Low Risk.

Profile G3-Low: Mild fatigue with transient bowel change and no bleeding. System triggered two small additive terms but no persistence multipliers. Output: Low Risk.

Transcript Set 2: Moderate-Risk Profiles **Profile G4-Moderate:** Occasional bleeding, moderate bowel change, and duration near two weeks. Interaction term for bleeding plus duration partially activated. Output crossed low threshold and remained below high threshold. Output: Moderate Risk.

Profile G5-Moderate: No bleeding but persistent bowel change and moderate fatigue over three weeks with positive family history. Additive context score raised final class into moderate band. Output: Moderate Risk.

Profile G6-Moderate: Recurrent symptom pattern with mixed severity but not all high-priority interactions active. Output stable in moderate range.

Transcript Set 3: High-Risk Profiles **Profile G7-High:** Persistent bleeding, persistent bowel change, prolonged duration, and fatigue. Multiple interaction bonuses activated. Score exceeded high-risk threshold with clear escalation guidance language.

Profile G8-High: Recurrent bleeding with long duration and positive family history. Even with moderate pain, combined persistence-context structure produced High Risk classification.

Profile G9-High: Persistent multi-symptom cluster across all key fields. Output remained High Risk across repeated runs, demonstrating deterministic consistency.

Transcript Set 4: Boundary Behavior Checks **Profile G10-Boundary:** Inputs designed to sit near low/moderate threshold. Small duration increase shifted class from Low to Moderate as expected.

Profile G11-Boundary: Moderate profile near upper threshold. Adding persistent bleeding shifted class to High.

Profile G12-Boundary: High profile with one feature reduced from persistent to occasional. Class decreased to Moderate, showing directional sensitivity in threshold mapping.

Appendix H: Deployment Readiness and Governance Checklist The following checklist guides progression from prototype to research-grade pilot.

Requirement Area	Current Status	Next Action
Data governance policy	Not complete	Define retention windows and access rules

Consent and disclosure language	Draft only	Formal review with legal and ethics advisors
Clinical advisory oversight	Not established	Recruit gastroenterology review panel
Retrospective validation dataset	Not available	Establish de-identified dataset partnership
Prospective study protocol	Not available	Draft and submit protocol for review
Threshold calibration plan	Prototype only	Refit on representative cohort data
Fairness and subgroup audits	Not run	Define subgroup metrics and acceptance thresholds
Security controls (production)	Not implemented	Add encryption, access control, and logs
Incident response plan	Not implemented	Create escalation and communication policy
Model versioning and rollback	Not implemented	Add deployment controls and rollback path
Human factors testing	Not run	Conduct usability sessions with target users
Clinician interface design	Not built	Prototype referral-support dashboard
EHR integration feasibility	Not assessed	Scope interoperability requirements
Regulatory pathway assessment	Not assessed	Seek regulatory counsel on classification category
Post-deployment monitoring plan	Not available	Define drift and quality monitoring standards

B Transparency, Ethics, and Reproducibility Statements

Ethics Statement This study uses secondary de-identified survey data and prototype-generated synthetic/derived labels for internal method development. No clinical intervention, no patient-facing diagnostic deployment, and no emergency decision workflow were conducted in this study.

Data Statement Analyses were performed on uploaded NHANES component files available in this project workspace [7, 8]. Integrated analysis artifacts and experiment outputs are stored under the project `data/` directory.

Code and Reproducibility Statement All experiment steps reported here were executed with deterministic scripts in this workspace, including preprocessing, splitting, and metric generation. The study reports internal validation only; external reproducibility should include independent cohort testing and clinically adjudicated outcome labels.

Conflict of Interest Statement The author reports development involvement in the described prototype system. This manuscript is intended as a methodological and feasibility report.

Clinical Use Disclaimer GastroLens is an informational research prototype and does not provide diagnosis, treatment, or emergency triage decisions.

References

[1] Cross E, Saunders B, Farmer AD, Prior JA. Diagnostic delay in adult inflammatory bowel disease: A systematic review. *Indian Journal of Gastroenterology*. 2023;42(1):40–52. doi:10.1007/s12664-022-01303-x.

[2] Sun J, Fang F, Olén O, et al. Long-term risk of inflammatory bowel disease after endoscopic biopsy with normal mucosa: A population-based, sibling-controlled cohort study in Sweden. *PLOS Medicine*. 2023;20(2):e1004185. doi:10.1371/journal.pmed.1004185.

[3] Gomollón F, Burisch J, Allez M, et al. ECCO-ESGAR-ESP-IBUS Guideline on Diagnostics and Monitoring of Patients with Inflammatory Bowel Disease: Part 1. *Journal of Crohn's and Colitis*. 2025. doi:10.1093/ecco-jcc/jjaf106.

[4] Stafford IS, Gosink MM, Mossotto E, Ennis S, Hauben M. A Systematic Review of Artificial Intelligence and Machine Learning Applications to Inflammatory Bowel Disease, with Practical Guidelines for Interpretation. *Inflammatory Bowel Diseases*. 2022;28(10):1573–1583. doi:10.1093/ibd/izac115.

[5] Majtner T, Brodersen JB, Herp J, et al. A deep learning framework for autonomous detection and classification of Crohn's disease lesions in the small bowel and colon with capsule endoscopy. *Endoscopy International Open*. 2021;9(9):E1361–E1370. doi:10.1055/a-1507-4980.

[6] Judge CS, Krewer F, O'Donnell MJ, et al. Multimodal Artificial Intelligence in Medicine. *Kidney360*. 2024;5(11):1771–1779. doi:10.34067/KID.0000000000000556.

[7] National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). National Health and Nutrition Examination Survey (NHANES): Survey Methods and Analytic Guidelines. Accessed March 28, 2026. Available at: `https://wwwn.cdc.gov/nchs/nhanes/analyticguidelines.aspx`.

[8] National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). National Health and Nutrition Examination Survey (NHANES) Data, Documentation, Codebooks, SAS Transport Files, and Frequency Tables. Accessed March 28, 2026. Available at: `https://wwwn.cdc.gov/nchs/nhanes/Default.aspx`.