Artificial Intelligence in Gastroenterology: Current Challenges, Critical Evidence, and Clinical Translation

Bahaaeldeen Ismail, MD

Disclosures

- No related financial disclosures

Need/Practice Gap

- AI capabilities in gastroenterology are expanding rapidly
- A significant gap exists between algorithm development and real-world clinical implementation
- **Clinical Need**: Practicing gastroenterologists need a structured framework to critically appraise emerging AI tools, distinguish true clinical utility from vendor hype, and safely integrate validated models into daily workflow.

Objectives

- Identify Clinical AI Capabilities in GI & Limitations
- Appraise Key Findings from DDW 2026 by analyzing selected high-impact AI data from the meeting
- Formulate a structured, physician-led approach to critically appraise, and mitigate clinical bias or liability when integrating new AI tools into routine GI practice.

Expected Outcome

- Critically Appraise Clinical AI in GI: Differentiate validated, high-yield GI AI applications from overhyped or under-tested algorithms
- Apply insights from DDW 2026 AI trials to clinical decision-making
- Establish clinical oversight protocols to safely adopt AI tools in GI practice

The AI Landscape in Healthcare and GI

- *Source: AI In Your Endoscopy Practice | Lines: 11–21*
- Adoption is real but uneven
  - 22% of healthcare organizations have deployed commercial AI products
  - 9% adoption across the broader U.S. economy — healthcare is ahead of most sectors
  - Most spending goes to ambient scribes, coding and billing, and workflow — not clinical decision support or diagnostics
- *Source: AI Colonoscopy and Adenoma Detection | Lines: 21–22*
- Colon polyp detection is the entry point for AI in endoscopy
  - 5 FDA-cleared CADe products, with widespread U.S. adoption since 2021

"Is AI truly accurate in real-world settings?"

- Algorithm performance ≠ real-world impact
- *Source: AI Colonoscopy and Adenoma Detection | Lines: 193–214, 276–319*
- Controlled RCTs show benefit
  - Meta-analysis of 44 RCTs: ↑ adenomas per colonoscopy and ↑ ADR with CADe
  - No difference in advanced colorectal neoplasia per colonoscopy; only a small ↑ in ACN detection rate
- Pragmatic real-world trials tell a different story
  - coloDETECT: improved detection, but operators knew they were in a trial
  - NAIAD (UK): ↑ ADR with CADe, returning to baseline when the device was switched off
  - Australia: ADR improved only among adopters — the lowest-ADR endoscopists did not use CADe

Why the RCT Benefit Fades in Practice

- *Source: AI Colonoscopy and Adenoma Detection | Lines: 209–218*
- Possible explanations for the RCT vs. real-world discrepancy
  - Alert fatigue from constant AI prompts during the procedure
  - False positives reducing the endoscopist's thrust to investigate
  - Hawthorne effect — observed operators behave better than usual
  - Selection bias in RCT populations
- Addressed by Barkin's triple-blinded design — see the design and results slides that follow

CADe Gains by Provider Experience Level

- *Source: Artificial Intelligence and Precision Education in Gastroenterology | Lines: 10–72*
- Design: multicentric retrospective study at two academic training centers (UAB + USA), 2022–2025; >4,800 average-risk screening colonoscopies
- Combined ADR/SDR gains with CADe
  - Attendings: almost +15% — the largest gain group
  - First-year fellows: +10.3%
  - Second-year fellows: moderate gains; third-year fellows: minimal (already near ceiling)
  - SDR increased across all experience levels: +2.5% to +7.5%
- Limitations: retrospective observational design, only two Alabama centers, variable attending supervision

Barkin Triple-Blinded RCT — Results

- *Source: AI Colonoscopy and Adenoma Detection | Lines: 188–320*
- 1,744 patients; the fully concealed design removes the Hawthorne effect entirely
- ITT analysis (CADe available in the room): no significant ADR difference
- Per-protocol analysis (CADe actually used): >5% absolute ADR improvement — significant
- Advanced lesion detection improved significantly in BOTH analyses — the key finding
- CADe utilization: only 57.8% overall (Center 1 ~70%, Center 2 <40%, Center 3 ~2%)

Will AI Replace Me or My Trainees? The De-Skilling Question

- Air France Flight 447 (2009): the autopilot disengaged at 37,000 feet and the pilots could not recover manually, having become dependent on automation
  - Pilots are now required to fly manually at intervals to maintain skills
- Implication for training: include withdrawals both with AND without AI
- Open question: is practicing without HD scopes "upskilling"? If trainees will never practice without AI, what baseline are we protecting?

Augmentation, Not Replacement: The Centaur Model

- *Source: AI In Your Endoscopy Practice | Lines: 279–298*
- The "Centaur model" from Formula One offers a framework
  - Human–AI collaboration for optimal performance
  - AI proposes constantly; the human frames and decides
  - The system learns from both machine and human feedback
  - Even 0.15 seconds matters for competitive advantage
- The evidence points toward augmentation — the Centaur outperforms either alone, but only when the human component is competent

"Should I trust AI tools? When do I override?"

- *Source: AI In Your Endoscopy Practice | Lines: 479–491*
- When the endoscopist disagrees with AI
  - All current products allow manual override of any component — a non-negotiable safety feature
- When AI is confident and the endoscopist is unsure
  - Confidence display (e.g., 50% vs. 85%) helps inform judgment
  - Human judgment ultimately prevails regardless of AI confidence
- Future research needs: how AI affects clinician uncertainty, and the dynamics of optimal human–AI collaboration

If AI Misses a Lesion, Who Is Responsible?

- "If AI misses a polyp or misclassifies a lesion, who's legally responsible?"
- *Source: AI In Your Endoscopy Practice | Lines: 243–245*
- The signing physician remains fully responsible for every word in an AI-assisted report
- AI is a tool, not a cosigner — the liability structure is unchanged
- Formally document AI involvement in report generation

"Is AI worth the investment for my practice/hospital?"

- *Source: AI Applications in Liver Diseases and Transplantation | Lines: 182–202*
- Economic evidence from the MASLD fibrosis screening study
  - AI + elastography delivered the highest QALYs and the lowest ICER, below the $100,000 threshold
  - Both semaglutide and resmetirom fell below the $100,000 ICER threshold when paired with AI screening
  - Results were robust under uncertainty, including at lower willingness-to-pay thresholds
- The full model and cost-effectiveness results appear in the Liver Disease section

COLONOSCOPY — The Most Mature AI Application in GI

Real-Time Polyp Detection (CADe): The Evidence Base

- *Source: AI In Your Endoscopy Practice | Lines: 42–65*
- The largest evidence base for any AI application in endoscopy: ~48 RCTs with >40,000 patients and 5 FDA-approved products
- Clinical impact
  - Normal detection rate ↑ 20%; adenoma miss rate ↓ approximately 50%
  - Adenomas per colonoscopy ↑ significantly, for all providers regardless of experience level
  - No significant impact on withdrawal time
- Caveats and limitations
  - Gains are driven primarily by diminutive polyps — is this clinically meaningful?
  - Real-world performance gains are smaller than RCTs demonstrate

Real-World AI Impact: TriNetX Study Design

- *Source: AI Colonoscopy and Adenoma Detection | Lines: 10–96*
- The largest real-world AI impact study in GI — federated EHR data from 67 healthcare organizations
  - Pre-AI era: January 2015 – December 2019 (~1,600,000 encounters)
  - AI era: January 2022 – June 2025 (~2,000,000 encounters)
  - 2020–2021 excluded for COVID disruption and early adoption transition
  - Adults aged 45–90 undergoing colonoscopy; CRC at or within 1 month of colonoscopy excluded
  - 1:1 propensity score matching (standardized mean difference <0.1) → ~1,500,000 per cohort

Outcome	Pre-AI Era	AI Era	Risk Ratio
Adenoma detection	1.8%	33.6%	**1.97**
Advanced adenoma detection	0.13%	0.19%	**1.48**
Interval CRC (6–36 months)	0.21%	0.11%	**0.53** (47% reduction)

TriNetX Study — Limitations

- *Source: AI Colonoscopy and Adenoma Detection | Lines: 10–96*
- An ecological comparison, not a direct AI vs. no-AI trial — other quality improvements occurred over the same period
  - Increased focus on withdrawal time and bowel prep quality; ADR feedback programs
- Reliance on ICD-10 diagnosis codes rather than direct pathologic review — less granular
- Many CADe systems with different sensitivities and integration workflows exist within the AI era
- Unmeasured confounders: endoscopist-level ADR, procedure quality metrics, sedation type, prep adequacy

Barkin Triple-Blinded RCT — Design

- *Source: AI Colonoscopy and Adenoma Detection | Lines: 188–320*
- Triple-blinded, fully concealed design
  - Patients unaware of trial participation; endoscopy staff unaware the trial existed; analysts blinded to allocation
  - Removes the Hawthorne effect from colonoscopy performance entirely
- Randomization by room availability rather than by use
  - Preserves authentic clinical behavior — endoscopists free to use CADe or not at their discretion
  - 1:1 randomization stratified by center, endoscopist, patient sex and age group
- Setting and population
  - Three Quebec high-volume academic centers over a 2-year period; patients aged 45–89 undergoing screening, surveillance or diagnostic colonoscopy
  - Operators: full-time board-certified gastroenterologists and general surgeons
  - Exclusions: emergency colonoscopy, known or suspected polyposis, known IBD
- Intervention: Medtronic GI Genius CADe system

CADx and Endoscopist Skill — Study Design

- *Source: AI Colonoscopy and Adenoma Detection | Lines: 322–406*
- The first longitudinal study of CADx impact on skill
  - 4-year prospective cohort (2021–2025), Montreal University Hospital Center
  - CADx system: CatEye — available throughout, with use at the endoscopist's discretion (not mandated)
- Cohort
  - 1,321 patients and 2,855 polyps (CADx-assisted 1,257 [44%]; unassisted 1,598 [56%])
  - Mean polyp size 7.25 mm; adenoma ~60%, hyperplastic ~20%, serrated ~6%
  - 30 endoscopists (17 trainees); 60% made at least one CADx-assisted diagnosis

CADx and Endoscopist Skill — Outcomes

- *Source: AI Colonoscopy and Adenoma Detection | Lines: 322–406*
- Primary outcome — the upskilling signal
  - CADx-assisted accuracy: ↑3.4% per month (statistically significant)
  - Unassisted accuracy: ↑1.1% per month (not statistically significant)
  - CADx-assisted curve rises steeply over 4 years; the unassisted curve plateaus around 65–70%
- Secondary outcomes — all metrics trending higher with CADx
  - Specificity +2.3% per month; positive predictive value +5% per month (both significant)
  - Unassisted cases showed none of these trends

CADx Upskilling — Mechanism and Limits

- *Source: AI Colonoscopy and Adenoma Detection | Lines: 322–406*
- Why CADx may upskill
  - Immediate feedback during polyp evaluation, versus pathology results roughly a month later
  - Reinforces correct pattern recognition and directs attention to cases of diagnostic uncertainty
  - Provides an objective second opinion in real time — rare without AI
- CADe vs. CADx — different behavioral effects
  - CADe: activated during withdrawal and passive for the endoscopist — may carry de-skilling risk
  - CADx: requires active participation (switch on, wash the polyp, center it) — this engagement may explain the upskilling effect
- Limitation: non-randomized design → significant selection bias risk

Why Lesions Are Missed — and How AI Can Teach

- *Source: AI In Your Endoscopy Practice | Lines: 104–116*
- Five reasons lesions are missed
  - Hard to see — subtle morphology, poor lighting, suboptimal angle
  - Technique, time and fatigue — withdrawal speed and experience level
  - Lack of mucosal exposure — poor bowel prep hides lesions
  - Not recognized — inattentional blindness: the lesion is visible but never consciously registered
  - Unfamiliar presentations — e.g., uncommon Barrett's lesions that general endoscopists rarely see
- *Source: AI-Driven Gastroenterology | Lines: 331–344*
- Trainee tandem study — a recognition failure, not a detection failure
  - Trainees missed 55% of serrated polyps and 31% of adenomatous polyps
  - 80% of misses: the polyp was on screen but not recognized; only 20% never appeared on screen
- AI as a teaching tool: flagged subtle lesions create a feedback loop for pattern learning

Multimodal LLMs for Polyp Analysis — Design and Learning

- *Source: AI Colonoscopy and Adenoma Detection, lines 409-500*
- Study by Kyung Kim, Seoul National University Hospital — MLLMs excel in radiology, pathology and reporting, but colonoscopy use remains limited
- Study design
  - Data: SUN dataset (>24,000 annotated images) and Real Colon dataset
  - Tasks: polyp detection and localization with bounding boxes
  - Benchmark: three internal medicine residents and three gastroenterologists
- In-context learning (Gemini 2.5 Pro)
  - Accuracy rose from 80% zero-shot to nearly 90% at 128 shots
  - Just 8 shots delivered meaningful gains — no large-scale training needed
  - Reached internal medicine resident level

Multimodal LLMs for Polyp Analysis — Results and Limits

- *Source: AI Colonoscopy and Adenoma Detection, lines 409-500*
- Fine-tuning (LLaMA 3.2 Vision + MedLLaMA)
  - Detection accuracy: 58% → 93%
  - Localization error (MAE): ~14% → 3%
  - Accuracy at IOU ≥0.5: 39% → 96%
- Versus human readers
  - Gastroenterologists remained most accurate, but the gap to Gemini was minimal
  - The model was more consistent than humans across repeated trials
- Implications and limits
  - Uses: endoscopy quality assurance and training where experts are scarce
  - Current work is on static images — video, 3D perception and size estimation are next

Machine Learning for Polyp Risk Stratification (VA Data)

- *Source: AI Colonoscopy and Adenoma Detection, lines 503-611*
- Study by Ethan Espinosa, VA Boston Health Care System — VA Corporate Data Warehouse, January 2010 to January 2025
- Cohort and features
  - ~408,000 patients screened to ~32,000 with two colonoscopies at least 8 months apart; 33 features across history, labs and pathology text
  - NLP with BioBERT converts free-text pathology into structured inputs: neoplastic polyp presence, count, location, subtype or grade
- Modeling and performance
  - Five models tested; AUC 0.71 to 0.72 across all, with 5-fold cross-validation variance ~0.01
- What drives predictions (SHAP)
  - Pathology findings dominate — 4 of the top 10 features; personal history and labs supply the other 6
  - Higher polyp counts, age near 60 and triglycerides near 100 mg/dL push risk up
- Application and limits: supports follow-up timing (7 versus 10 years); 90% data exclusion risks selection bias

CATe for Barrett's Neoplasia — Prospective Multicenter Results

- *Source: AI Meets the Esophagus | Lines: 10–52*
- Design: prospective multicenter study at 3 Dutch expert centers; 200 Barrett's patients; stepwise expert assessment as ground truth
- Results
  - 82 patients had visible lesions; 73/82 detected → sensitivity 89%
  - 74 false positive detections in 58 of the 118 lesion-free patients → 0.6 per patient (about one extra biopsy every two patients, versus 16 random biopsies)
  - Most detected lesions were neoplastic (HGD or cancer); ~1/3 non-neoplastic but still true endoscopic detections
  - Missed lesions were flat, small or near gastric folds — subtle features remain challenging even for expert-level AI

CATe — Why the Endpoint Choice Matters

- *Source: AI Meets the Esophagus | Lines: 10–52*
- Histology-proven neoplasia is a problematic endpoint
  - Low prevalence in community settings requires extremely large sample sizes
  - Biopsy sampling may be inadequate or miss the lesion entirely
  - Inter-observer variability among pathologists persists
- Recommendation: shift to endoscopic endpoints (visible lesion detection), aligned with the purpose of CADe/CATe systems
- A large Chinese RCT (~30,000 patients) showed only minimal AI benefit when evaluated against histology — the endpoint choice matters

NLP and LLMs for Barrett's Surveillance Scheduling

- *Source: AI-Driven Gastroenterology | Lines: 66–71, 405–408*
- NLP extracts data from endoscopy and pathology reports and sets surveillance intervals without human review — ~90% accuracy versus guidelines, feeding directly into the EMR
- *Source: AI Meets the Esophagus | Lines: 149–180*
- LLM-based system with a rules-based backend
  - Extracts surveillance-relevant Barrett's variables, dysplasia grade and segment length
  - A rules-based engine assigns guideline-concordant intervals — reducing hallucination risk and keeping benchmarking transparent
- Validation
  - Internal: excellent per-variable extraction and strong guideline concordance, with tailoring by segment length and longitudinal tracking through eradication therapy
  - External: multicenter veterans cohort (pathology data only) maintained performance and captured edge cases regex/NLP miss
- Future: EHR integration through a FHIR interface; testing for trust, usability and effectiveness

Blood-Based Biomarkers for Esophageal Cancer: REG4 — Discovery

- *Source: AI Meets the Esophagus | Lines: 263–302*
- Plasma proteome via Olink proximity extension assay — ~3,000 proteins quantified per sample
- UK Biobank proteomic cohort (~50,000 healthy participants)
  - Baseline blood draw with years of follow-up for recorded disease development, including esophageal cancer
  - Cox proportional hazards model; 50 random seeds, 50/50 case-control splits (3 controls per case)
  - Blood drawn 5–6 years before clinical diagnosis still achieved AUC ≈ 0.90
  - A parsimonious 4-protein model plus clinical covariates also reached AUC ≈ 0.90
- Key driver protein: REG4 — most frequently included across the 50 models

REG4 — Independent Validation and Performance

- *Source: AI Meets the Esophagus | Lines: 263–302*
- Independent validation cohort: 171 symptomatic reflux patients with same-day blood draw and endoscopy
  - REG4 concentration rises progressively as disease progresses, with significant trend testing
- Classification performance by stage
  - All cancer versus controls: AUC 0.88
  - T1b and beyond: AUC 0.96
  - High-risk (dysplasia + cancer) versus symptomatic reflux and non-dysplastic Barrett's: AUC ≈ 0.80 with REG4 alone
- Biological validation: Visium transcriptomic data and immunohistochemistry confirm REG4 expression

AI for Predicting GERD Treatment Response

- *Source: AI Meets the Esophagus | Lines: 182–217*
- Problem: responses to standard acid suppression vary widely and current prediction tools are limited
- Design: multi-center real-world study (Guangdong, China); five AI models built on clinical data plus three objective esophageal function tests
- Efficacy prediction
  - Full model (153 parameters, internal 5-fold cross-validation): XGBoost performed best
  - Internal validation (678 patients): AUC 0.67; external validation (114 patients): AUC 0.63; average accuracy 0.65–0.70
  - Simplified 10-parameter models reach ~70% accuracy with or without functional test results
- Medication optimization: random forest regression, mean absolute error 3.9 points on a 40-point symptom scale; acid suppressant monotherapy was the most common optimal regimen, then acid suppressant plus prokinetic
- Conclusion: simplified, symptom-based models retain good predictive value and can recommend optimal regimens

Machine Learning for Variceal Hemorrhage Outcomes — Design

- *Source: AI Meets the Esophagus | Lines: 53–107*
- Why new scores are needed
  - MELD ignores hemodynamics and endoscopy; CHIPS is categorical and imprecise; Rockall/Glasgow are pre-endoscopic and ignore cirrhosis and endoscopy data
- Design: retrospective MIMIC-IV cohort (Beth Israel, 2008–2022); acute variceal hemorrhage confirmed by EGD text (banding or cryotherapy injection)
- Predictor variables
  - Demographics and comorbidities; first 24-hour clinical data (BP, HR, MAP, labs); transfusion requirements; vasoactive medication use
  - NLP-extracted endoscopic features: variceal size and location, active spurting, red wale sign, portal hypertensive gastropathy grading

Variceal Hemorrhage Model — Performance and Clinical Use

- *Source: AI Meets the Esophagus | Lines: 53–107*
- Model performance (XGBoost)
  - 6-week mortality: AUROC 0.89, AUPRC 0.68 versus MELD-Na 0.78 and CHIPS/Rockall ≈ 0.7
  - 6-week rebleeding: AUROC 0.80 versus MELD-Na 0.7
- Top predictors of mortality (SHAP): lactate, creatinine, hemodynamic instability, active spurting on EGD, esophageal varices, early transfusion, vasoactive drug use
- Clinical decision pathway
  - Low risk → standard follow-up; intermediate → hepatology consult and early TIPS consideration; high risk → ICU, early TIPS, prophylaxis
- Limitations: external validation needed; ICD misclassification; no pre-admission outpatient data; prospective trial required

AI for Supra-Gastric Belching Detection

- *Source: AI Meets the Esophagus | Lines: 109–146*
- Background: events last tenths of a second and occur sporadically, so manual identification is tedious — about 30 minutes per study
- Design: 15 GERD patients; pH-impedance data (~90,000 lines per study) exported, meals removed, 2-second windows filtered at ≥500 ohm rise above baseline; a neural network classifies true versus false events against blinded manual marking
- Results
  - Sensitivity ≈80% overall, rising significantly when studies with 0–1 events are excluded
  - Mean 2 false positives per study (highest 8); analysis completed in seconds
  - The algorithm identified events missed by humans — added clinical value
- Clinical validation: normal and excessive counts agreed with manual review, including cases near the >13-event threshold

Transformer AI for FLIP Panometry Classification — Design

- *Source: Esophageal Motility Disorders | Lines: 166–220; DDW 2024 Esophageal Motility | Lines: 145–193*
- Background: FLIP classification was refined to version 2.0 (Dallas consensus) while prior CNN models used v1.0; inter-rater reliability among specialists is only moderate-to-good
- Transformer advantages over CNNs
  - Self-attention focuses on the relevant parts of the input rather than local features
  - Captures spatial and temporal dependencies — well suited to longitudinal physiologic data
- Novel cost-sensitive penalty matrix
  - 0 cost for correct classifications; 100 units for critical misclassifications (e.g., achalasia → normal); 10–50 units for diagnostically adjacent errors
  - Penalizes errors in proportion to patient impact — a shift from treating all mistakes equally
- Cohort: n = 984 (2013–2024); mean age 53, 59% female; normal 36%, abnormal 25%, achalasia 39%

FLIP Transformer — Performance and Clinical Concordance

- *Source: Esophageal Motility Disorders | Lines: 166–220; DDW 2024 Esophageal Motility | Lines: 145–193*
- Model performance
  - Overall accuracy 89% (3-class); binary accuracy 94%
  - Normal: sensitivity 89%, specificity 97%; abnormal: 85% and 91%; achalasia: 93% and 96%
  - No achalasia classified as normal and no normal classified as achalasia — zero critical errors
- Cost matrix impact: eliminated all high-priority misclassifications and reduced moderate ones; remaining errors carried a penalty ≤25
- Clinical concordance
  - FLIP v2.0 normal has 99% NPV for achalasia/conclusive EGJOO
  - Non-spastic obstruction on FLIP has 91% PPV for achalasia/conclusive EGJOO
  - AI model performance matches expert human concordance rates

PANCREAS & BILIARY

EUS AI for Pancreatic Mass Characterization

- *Source: AI-Driven Gastroenterology | Lines: 84–117*
- Background: EUS-FNA sensitivity for lesions <10 mm is 82% with 91% accuracy; an opportunity exists for quality metrics such as percentage of pancreas visualized
- Key studies
  - Mayo (Neil/Levy): 583 patients with both videos and images; AI outperformed physicians in sensitivity while maintaining specificity, and identified autoimmune pancreatitis via bile duct wall thickening
  - Japan (772 patients): single-center retrospective model using data augmentation to correct an 80% cancer imbalance; high sensitivity, but external validation is needed
  - China: multimodality ensemble combining labs, history, radiology and EUS enhanced cancer detection
  - Tandem study: 36 EUS videos with many endosonographers — reviewer accuracy 77% with AI
  - Real-time EUS-AI detects and segments pancreatic mass lesions comparably to expert annotation

AI Impact on Endosonographer Performance — Study Design

- *Source: Artificial Intelligence Applications in Pancreatic Diseases | Lines: 213–282*
- Context: early pancreatic cancer detection remains challenging and EUS miss rates for subtle lesions are high; AI may support real-time interpretation
- Objective: evaluate MyCompanion (iGlobe) for lesion detection accuracy and diagnostic confidence in solid pancreatic tumors
- Design: randomized controlled tandem video reader study
  - 38 curated, histology-confirmed 2-minute EUS clips (true positive and true negative), not used in AI training
  - 57 endosonographers of varying experience across 3 sequential experiments of increasing difficulty
  - Randomized first reading with or without AI; bounding-box output; single-pass review; confidence rated on a 3-point Likert scale

AI Impact on Endosonographer Performance — Results

- *Source: Artificial Intelligence Applications in Pancreatic Diseases | Lines: 213–282*
- Overall detection accuracy: 70.8% without AI versus 77.8% with AI (+7% absolute)
- By experience level
  - Novice and experienced readers both benefited in the harder experiments 2–3
  - No significant benefit in experiment 1 (easy cases); unexpectedly, even expert endosonographers improved
- Diagnostic confidence improved significantly with AI assistance
- Benefit was most pronounced for subtle, small (<2 cm) lesions in difficult cases
- Limitations: benchtop study — scope handling and rescanning are not captured; clinical studies are ongoing

Deep Learning for Pancreatic Cancer on CT: The Spark Model — Design

- *Source: Artificial Intelligence Applications in Pancreatic Diseases | Lines: 284–365*
- Clinical problem: ~38% of stage IV patients had subtle CT features on prior imaging; stage I survival exceeds 83% — a recognition failure, not a detection failure
- Objective: recognize subtle early CT features and deliver an accessible web-based tool
- Data and model
  - ~2,000 CT images from multiple institutions worldwide, classified by Yale and Augusta University radiologists
  - EfficientNetB4 (~6–8 million parameters) chosen over 100–150M-parameter alternatives to avoid overfitting and maintain reproducibility
  - 60% training, 20% internal validation, 20% external validation; images segmented to the pancreas with augmentation across angles and dimensions

Spark Model — Performance and Deployment

- *Source: Artificial Intelligence Applications in Pancreatic Diseases | Lines: 284–365*
- Performance
  - Internal validation accuracy >99%, with high AUROC, F1 and F2 scores
  - External validation via the website dipped as expected but maintained high accuracy, with minimal false positives and negatives
- Computational efficiency: 0.7 seconds for a single image; under 60 seconds for 1,000 images
- Deployment: hosted on a website with no high-performance computing or GPU required — accessible on standard computers globally

Multi-Organ Radiomics for Early PDAC Prediction — Design

- *Source: Artificial Intelligence Applications in Pancreatic Diseases | Lines: 367–478*
- Background: no general-population screening exists (PDAC incidence ~13/100,000); high-risk subgroups have ~1% annual incidence
  - Prior Mayo study (2022): ML on CTs 3–36 months before diagnosis reached AUC 0.98 versus 0.66 for human readers, but ~40% had pancreatic abnormalities on radiologist review
- Objective: an imaging biomarker combining multi-organ radiomics with body composition to identify a risk-enriched cohort
- Design: retrospective case-control — confirmed PDAC with incidental CTs 2–60 months before diagnosis; controls matched 3:1 on race, sex, CT year and contrast status
- Features: ~75 radiomics features per organ (pancreas, liver, spleen, kidney) plus L3 body composition (skeletal muscle and adipose area and attenuation)
  - Rationale: muscle wasting is detectable up to 18 months and adipose change up to 6 months before diagnosis
- Pipeline: batch harmonization → feature selection (Mann-Whitney $p<0.1$) → regularization → normalization → 5 ML classifiers; 80% training/validation, 20% test

Multi-Organ Radiomics — Results and Implications

- *Source: Artificial Intelligence Applications in Pancreatic Diseases | Lines: 367–478*
- Risk scores rise significantly as patients approach diagnosis and differ from controls across all pre-diagnostic windows
- Feature importance by window
  - 2–5 months: pancreas shape dominates — top feature is pancreas spherical disproportion
  - 6–24 months: mixed features — pancreas shape and texture, liver diameter, adipose tissue
  - 25–60 months: body composition and liver features — subcutaneous fat attenuation, visceral adipose tissue
- No performance difference by metastatic status or tumor site — the signature reflects systemic change, not just local tumor
- Implication: may identify candidates for intensified screening 6–24 months before diagnosis, even when the pancreas appears normal

Model	AUROC
Pancreas radiomics only	0.69
Liver radiomics only	0.67
Body composition only	**0.57**
Pancreas + Liver + Body composition (multi-organ)	**0.72**

Pre-Diagnostic Window	AUROC
2–5 months	**0.79** (highest)
6–24 months	**0.71**
25–60 months	**0.68** (lowest)

Biliary Stricture Assessment: AI vs. Tissue Sampling

- *Source: AI-Driven Gastroenterology | Lines: 139–178*
- Problem: visual examination alone has poor inter-observer agreement — experts are sensitive but cannot agree on individual cases
- Mayo model: 10,000 annotated images; 900-frame (~30 second) analysis; accuracy ~91%, sensitivity 93%, specificity lower than brush or forceps
- Multicenter validation: 61 cholangioscopy videos from 3 sites; AI sensitivity exceeded brush cytology and forceps biopsy, with much higher overall accuracy than tissue sampling
- SMART trial: real-time cholangioscopy AI outperformed combined brush cytology and transpapillary biopsy; experts performed no better than untrained trainees on the same videos
- Interface: running benign versus malignant score with a waterfall plot for cancer probability
- Future needs: external validation on large datasets, continuous feedback loops, real-time assessment and clearer regulatory pathways

LIVER DISEASE

AI Screening for Advanced Fibrosis: The Problem and FibroX

- *Source: AI Applications in Liver Diseases and Transplantation | Lines: 140–227*
- MASH/MASLD burden
  - ~1/3 of adults globally affected; 50 million US adults have MASH and ~7 million are already at F2+ fibrosis
  - Leading driver of cirrhosis, HCC and death; ~$76 billion in direct US healthcare costs; most advanced fibrosis remains undiagnosed
- Current pathway limitations: sequential FIB4 → elastography, with imperfect sensitivity and low real-world completion rates
- FibroX — an explainable LGBoost model developed at Yale
  - >20 clinical inputs from EHR records; automatically flags high-risk patients and explains why
  - AUC 97% versus FIB4 at 62% — the largest performance gap reported in MASLD literature

FibroX Cost-Effectiveness — Model and Methods

- *Source: AI Applications in Liver Diseases and Transplantation | Lines: 140–227*
- Markov state-transition model with 13 health states: fibrosis stages → cirrhosis → HCC → transplant → death, entering at MASLD/early fibrosis with annual transitions
- Four strategies compared
  - AI followed by elastography
  - FIB4 followed by elastography
  - Elastography alone
  - No screening
- Treatments modeled: semaglutide (GLP-1 agonist) and resmetirom (THR-β agonist), both FDA-approved for NASH
- Economic outcomes: total cost, QALYs, ICER and net monetary benefit, against a $50,000–$100,000 willingness-to-pay threshold

FibroX Cost-Effectiveness — Results and Limitations

- *Source: AI Applications in Liver Diseases and Transplantation | Lines: 140–227*
- Base case (10-year horizon): AI + elastography delivered the highest QALYs and lowest ICER, below the $100,000 threshold; other strategies had significantly higher ICERs
- Cost-effectiveness curves: AI + elastography remained optimal at lower willingness-to-pay thresholds and was robust under uncertainty
- Semaglutide dominates resmetirom — higher probability of being most effective at a lower threshold, showing that screening value is tied to treatment economics
- Why AI + elastography wins: better identification of high-risk patients and earlier treatment, with long-term benefit offsetting higher upfront screening cost
- Real-world implication: direct EHR integration enables scalable case finding — a high-value population health strategy for payers and health systems
- Limitations: model-based inputs from published literature and trial data; payer perspective only; real-world adherence and implementation costs not fully captured

HCC Treatment Allocation — Model Development

- *Source: AI Applications in Liver Diseases and Transplantation | Lines: 66–138*
- Problem: tumor board allocation combines oncologic knowledge with physician experience, but population-level statistics leave individual prognosis uncertain
- Data
  - ~16,000 patients from two large Italian registries; 80/20 training-validation split with bootstrap internal validation
  - External validation at the University of Tokyo; final model ~7,000 patients across all treatment types
  - Treatments: resection ~4,000; thermoablation ~1,487; chemoembolization ~1,164; systemic ~181; primary transplantation ~100
- Architecture: LASSO variable selection with a random forest (1,000 bootstrap validations), chosen over XGBoost, SVM and neural networks for the best performance-to-complexity ratio
- Performance: internal 1-year AUC 83% and 5-year 76%; external Japanese cohort 5-year AUC 72%

HCC Treatment Allocation — Reallocation Findings

- *Source: AI Applications in Liver Diseases and Transplantation | Lines: 66–138*
- The model predicts survival under each treatment and recommends the option with the highest predicted survival
- Reallocation findings
  - Liver transplantation: delivered in only 1.5% of cases versus best choice at 5 years in up to 22% — a 14-fold gap
  - Embolization: delivered in 16% versus best strategy in only 1.7% — a 9-fold overuse
  - 27% of patients completely reallocated at 1 year; 55% at 5 years
- Median predicted survival gain if recommendations are followed: +2% at 1 year and +24% at 5 years
- Clinical application: a web-based tool taking tumor board variables and explaining why a treatment is recommended
- Limitations: oncologic knowledge only — technical feasibility and local expertise still apply, and the final decision remains with the clinical team

Explainable AI for Cirrhosis Survival — Design

- *Source: AI Applications in Liver Diseases and Transplantation | Lines: 331–449*
- Problem: ~1 million US cirrhosis hospitalizations annually; 20% die within 90 days and >50% within a year once decompensated — a narrow window for ICU escalation, transplant evaluation or goals of care
- Limitations of existing scores
  - MELD AUC ~0.78; Child-Pugh is a static snapshot; CLIF-C ACLF is primarily for ICU patients
  - All rely exclusively on labs — none incorporate MAP, lactate, vasopressor requirement or multi-organ dysfunction signals
- Data: MIMIC-IV (2008–2022, Beth Israel); ~364,627 hospitalizations → ~11,000 patients after cirrhosis coding, index-admission and exclusion criteria
- Features: demographics, comorbidities, admission vitals, first 24-hour labs and organ support
- Outcome: 90-day transplant-free survival; XGBoost with Bayesian tuning (100 trials, 5-fold stratified CV), benchmarked against logistic regression and MELD

Explainable AI for Cirrhosis Survival — Results

- *Source: AI Applications in Liver Diseases and Transplantation | Lines: 331–449*
- XGBoost AUC 0.89 versus MELD 0.78 — a +14% discrimination improvement
- Decision curve analysis: XGBoost dominates all comparators, with a 15–35% gap over MELD in the clinically meaningful threshold zone
- Key predictors (SHAP): total bilirubin (strongest), sodium, INR, MAP, serum lactate and vasopressor use
- Subgroups: alcohol-related cirrhosis AUC ~0.90 and MELD 8+ ~0.86; NASH, viral hepatitis, ICU and ward subgroups all maintained performance
- Clinical utility: better first-24-hour triage, transplant referral timing and discharge planning, with patient-level SHAP explanations that build clinical trust
- Limitations and next steps: single-center retrospective data with a surrogate endpoint and no prospective test; external validation and real-time EMR red-flag alerts planned

AI Chatbot for Covert Hepatic Encephalopathy — Rationale and Design

- *Source: AI Applications in Liver Diseases and Transplantation | Lines: 229–328*
- Background: covert HE lacks findings on routine exam but is detectable with validated tools such as PHES; prevalence ~41% in cirrhosis and up to 60% in Child C
  - Testing is time-consuming and ~40% of AASLD providers have never performed dedicated CHE testing
- Objective: a text-based chatbot for covert HE detection, deployable at scale through patient portals such as Epic MyChart
- Design: pilot of 20 outpatients with chronic liver disease — 10 PHES-positive cases and 10 controls; excluded overt HE, psychotropics including opiates, and non-English speakers
- Platform: Laversa, a HIPAA and PHI compliant LLM using retrieval augmented generation over ~30 society guideline documents, written for a 6th-grade reading level
- Sessions supervised by a research coordinator for technical issues only, with no topic or length constraints — mimicking natural portal use

AI Chatbot for Covert HE — Analysis, Results and Limits

- *Source: AI Applications in Liver Diseases and Transplantation | Lines: 229–328*
- Pipeline: chat transcripts → OpenAI text embeddings (>3,000 dimensions) → PCA to 16 components explaining ~95% of variance; four models tested (k-means, decision tree, CNN, logistic regression)
- Cohort: median age 67, >50% male, case group median MELD 9.5; MASLD/MASH the most common etiology, then alcohol-related liver disease
- Results
  - K-means clustering performed best, with test characteristics ~0.60 — better than chance
  - Supervised models varied widely and performed poorly in leave-one-out cross-validation given the small sample and high dimensionality
  - Exploratory: in 6 returning patients, principal component 3 of between-visit change correlated negatively with final PHES score
- Limitations: selection bias toward older in-person visitors, difficulty typing, variable transcript richness and an observation effect from coordinator presence
- Future: validation in larger real-world datasets, portal deployment analyzing existing messages, and automated flagging of clinicians over time

Education/Training

GIESAP: An Adaptive Self-Assessment Tool for GI Fellows

- *Source: Artificial Intelligence and Precision Education in Gastroenterology | Lines: 146–254*
- Program structure — GI Endoscopy Self-Assessment Program
  - Streams of 5 weeks with 2–3 case-based questions every other day, each stream with a unique question bank
  - Questions are training-level specific (PGY-1 to PGY-3), developed from a validated surgery education program
  - Immediate answer reveal with detailed explanations, high-yield diagnostic information and analysis of incorrect options
  - ASGE partnership supplies high-quality endoscopic images and videos within answer explanations
- Proficiency is defined as answering a question correctly twice in a row
- Adult learning theory: the Ebbinghaus forgetting curve (50–75% lost within 24 hours) countered by spaced repetition

GIESAP — Pilot and Multicenter Results

- *Source: Artificial Intelligence and Precision Education in Gastroenterology | Lines: 146–254*
- Pilot (single center): 15 GI fellows across all training years in streams 1 and 2; an iterative neural network adjusted topic, timing and frequency in real time
  - Clear proficiency improvement across all training years; higher engagement predicted greater gains
  - Largest gains in the lowest-scoring topics: colorectal, esophagus, bariatric endoscopy and stomach
- Expansion (multi-center): >15 medical centers and >190 fellows across streams 1 through 6, with unchanged methodology
  - Improvements across all training years, durable and replicable in a diverse national sample
  - Heat map: improvement in every topic — liver doubled, small bowel and colorectal rose ~20–30%
- Covers topics often absent from traditional board prep: bariatric endoscopy, endohepatology, patient prep and professional development

GIESAP — Program Value and What Comes Next

- *Source: Artificial Intelligence and Precision Education in Gastroenterology | Lines: 146–254*
- Benefits: rapid integration of new guidelines and access to material not yet in traditional board prep — critical for continued clinical success
- Analytics for program leadership: monitoring mastery, tracking knowledge acquisition and identifying curriculum weak spots for data-driven change
- Next: moving from predictive to generative AI — a case-based generative platform that reassembles existing material into scenarios, questions and video-backed teaching on demand
- Feature set: scenario-based testing, science-backed spaced repetition, leaderboards and recognition, readiness analytics and coaching dashboards

LLMs in Clinical Decision Support — Study Design

- *Source: Artificial Intelligence and Precision Education in Gastroenterology | Lines: 257–377*
- Background: LLMs are posited to support clinician information needs, but real-world effects on practice and education remain unclear
  - Prior work on GutGPT found higher effort expectancy (ease of use) but no difference in intention, and did not examine knowledge or confidence
- Objectives: the effect of AI decision support with versus without GutGPT on knowledge, confidence and adoption propensity for upper GI bleeding — across medical students through senior consultants
- Tool: an ML risk dashboard (patient-specific 0–1 risk score, similar-patient counts, contributing features) with GutGPT layered on top, trained on UGI bleeding guidelines with cited references
- Design: randomized to dashboard + GutGPT or dashboard alone; 5 simulated UGI bleeding scenarios in a realistic emergency ward with EMR access and a simulated patient; surveys before, mid and after, plus interviews
- Enrollment: 80 of a target 100 — 42 physicians and 38 students; 42 in the GutGPT arm and 38 dashboard-only

LLMs in Clinical Decision Support — The Experience Level Effect

- *Source: Artificial Intelligence and Precision Education in Gastroenterology | Lines: 257–377*
- Knowledge and confidence rose in both arms; the knowledge gain was larger in the GutGPT arm but not statistically significant, and was driven mostly by medical students
- Adoption propensity
  - Use intention rose marginally only among students in the dashboard-only arm — no increase among GutGPT students or physicians
  - Explainability perception rose for dashboard-only students and for GutGPT-arm physicians
- Why students preferred the dashboard: 0–1 risk scores are more translatable, while GutGPT caused cognitive overload and prompt formulation was difficult with less experience
- Why physicians preferred GutGPT: experience allowed them to verify LLM responses, which confirmed their initial decisions
- Takeaway: the same tool helps or hinders depending on user experience — design AI interactions for the audience

Selection and Implementation of New AI Tools

Bias and Generalizability — The Core Risks

- *Source: AI In Your Endoscopy Practice | Lines: 34–35, 247–251*
- Biased training data means the algorithm reinforces existing disparities; performance degrades under data shift to a different population
- Many systems are trained at a few beta testing sites — monitoring in your own setting is essential
- *Source: AI-Driven Gastroenterology | Lines: 371–380*
- Geographic and regional training data concerns
  - Barrett's AI developed where dysplasia is common may not generalize to regions with different prevalence
  - FDA approval requires training in diverse settings including community centers, but the "substantially equivalent" pathway is less rigorous — a potential gap
- *Source: Artificial Intelligence Applications in Pancreatic Diseases | Lines: 489–496*
- Single-center retrospective designs with small samples are the most common limitation cited across DDW presentations — diverse datasets across ethnicities, institutions and scanner types are needed

External Validation Done Right — Two Positive Examples

- *Source: AI Colonoscopy and Adenoma Detection | Lines: 142–156*
- EfficientNet B7 histology model
  - External validation via website confirmed generalization — error rate only 0.87% on 2,000 unseen images
  - Lightweight architecture (66M parameters versus 100–150M alternatives) helps overcome overfitting
- *Source: Artificial Intelligence Applications in Pancreatic Diseases | Lines: 164–172*
- Pancreatic cyst AI — multi-continental validation
  - 40 EUS procedures from expert endoscopists across 3 continents; >21,000 images; 5-fold cross-validation to avoid overfitting
  - Interoperable across EUS processor brands through patented layer adaptation

Ask Vendors About Validation — and Verify the Answers

- *Source: AI In Your Endoscopy Practice | Lines: 392–417*
- Core vendor questions
  - Problem fit — does the solution address a real problem you have?
  - Scalability — can it layer on additional tools, or does it lock you in?
  - Compatibility — does it integrate with your equipment and EHR, ideally plug-and-play (HDMI, SD ports)?
  - Regulatory status — FDA approved, and for which indications?
  - Total cost of ownership — device, maintenance, contracts and EHR integration fees
  - Data ownership — hospital, physician or third party? Critical for future research and revenue
- Technical evaluation
  - Edge versus cloud deployment — latency, privacy and reliability trade-offs
  - Latency acceptable for real-time CADe/CADx use — delays defeat the purpose
  - Evidence pathway — De Novo versus 510(k) "substantially equivalent"; and the liability framework when AI fails
  - Interface design — bounding box versus segmentation mask, probability display, foot pedal or voice control; this determines whether the tool is actually used

Red Flags — What to Watch Out For

Red Flag	What It Means	Action
Single-center training data only	May not generalize to your population	Ask for external validation data
No external validation published	Performance claims are unverified in real settings	Proceed with caution; pilot before full deployment
"Black box" with no explainability	You cannot audit why it makes decisions	Require SHAP/LIME or similar interpretability tools
Vendor won't clarify data ownership/privacy	Your patient data may be monetized without consent	Get it in writing; review HIPAA compliance directly
Claims 99%+ accuracy on internal data only	Expected — external performance is always lower	Ask for external validation performance specifically
No human override capability	Unsafe and non-compliant with liability standards	Do not purchase — this is a deal-breaker
"Substantially equivalent" FDA pathway only	Less rigorous validation than De Novo approval	Ask what the predicate device was and how similar your use case is

Build a System, Not Just Buy a Device — The Four-Layer Architecture

- *Source: AI In Your Endoscopy Practice | Lines: 312–343, 367–389*
- Paradigm shift: from "What device should I buy?" to "What system should I build?"
- Hub design: procedure room → AI triage hub → smart scheduling → pathology integration loop → data governance layer → quality registry → continuous quality improvement loop
- The four layers of the modern endoscopy suite: real-time vision, documentation, pathology and administration
- Integration strategy: don't ask which product you need — ask which layer needs strengthening

Layer	Applications	Maturity Level
Layer 1: Real-Time Vision	CADe (polyp detection), CADx (optical diagnosis), Barrett's detection, quality metrics capture	Most mature — FDA-approved products available
Layer 2: Documentation & Reporting	Ambient clinical scribes, voice-to-report systems, image landmark auto-detection	Growing — several products in market
Layer 3: Tissue & Pathology	Digital pathology integration, multi-modal data pipeline development	Emerging — reimbursement implications for ASC/pathology ownership
Layer 4: Administrative AI	Scheduling optimization, workflow automation, revenue cycle management	Early stage — significant potential for efficiency gains

What Clinicians Should Do Today — Seven Actionable Steps

- Audit your current practice — identify which layer is your bottleneck: real-time vision, documentation, pathology or administration
- Evaluate vendors systematically — problem fit first, then scalability, compatibility, regulatory status, cost of ownership and data ownership
- Demand external validation — performance data from settings that resemble your own population, equipment and conditions
- Measure with AND without AI — maintain parallel tracking of outcomes in both conditions
- Train the entire team — nurses, technicians and physicians, with feedback loops so frontline staff report issues
- Establish governance from day one — monitoring, human override, regular validation and audit trails
- Think in systems, not products — strengthen your bottleneck layer, then build the next layer on top

Final Take-Home Messages (1 of 2)

- AI is here and growing — 22% of healthcare organizations have deployed AI; the question has shifted from whether AI arrives to how to integrate it safely
- Real-world performance < trial performance — the triple-blinded RCT showed no ADR improvement when CADe was merely available; always demand external validation in settings like yours
- Adoption is the bottleneck — only 57.8% utilization in the Barkin RCT; having AI available is not having it used, so address workflow and culture, not just technology
- CADx may upskill; CADe may deskill — passive detection carries de-skilling risk, while active diagnosis with immediate feedback appears to build skill over time
- You remain legally responsible — AI is a tool, not a cosigner; document its use, maintain override capability and establish audit trails

Final Take-Home Messages (2 of 2)

- Build systems, not just devices — think in layers: vision → documentation → pathology → administration
- The Centaur wins — human–AI collaboration outperforms either alone, but only when the human component is competent and engaged; invest in training alongside technology
- Learn from every case — build continuous feedback loops, borrowing from aviation, radiology and DevOps
- AI sees what humans cannot — fibrosis on MRI, subtle CT features in pancreatic cancer, impedance patterns in functional dysphagia; this extends human capability rather than replacing it
- Experience level matters — structured dashboards help less experienced users while experienced clinicians leverage LLMs effectively; design AI interactions for your audience