Article:
Dhali, A., Maity, R., Biswas, J. et al. (2026) Artificial intelligence-augmented small bowel capsule endoscopy for coeliac disease: a literature review on accuracy, workflow, and safety. Translational Gastroenterology and Hepatology, 11. 27. ISSN: 2224-476X

https://doi.org/10.21037/tgh-25-128

Review Article

Artificial intelligence-augmented small bowel capsule endoscopy for coeliac disease: a literature review on accuracy, workflow, and safety

Arkadeep Dhali[1,2#^], Rick Maity[3#], Jyotirmoy Biswas[4], Alexander Hann[5], Reena Sidhu[1,2], David Surendran Sanders[1,2]

[1]Academic Unit of Gastroenterology, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK; [2]School of Medicine and Population Health, University of Sheffield, Sheffield, UK; [3]School of Medical Science and Technology, Indian Institute of Technology Kharagpur, Kharagpur, India; [4]College of Medicine and Sagore Dutta Hospital, Kolkata, India; [5]Department of Internal Medicine II, Gastroenterology, University Hospital Würzburg, Würzburg, Germany

Contributions: (I) Conception and design: R Maity; (II) Administrative support: R Maity; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: A Dhali, J Biswas; (V) Data analysis and interpretation: A Dhali, J Biswas, A Hann, R Sidhu, DS Sanders; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

Correspondence to: Arkadeep Dhali, MBBS, MPH, FHEA. School of Medicine and Population Health, University of Sheffield, Sheffield, UK; Academic Unit of Gastroenterology, Sheffield Teaching Hospitals NHS Foundation Trust, Herries Road, Sheffield S5 7AU, UK. Email: arkadipdhali@gmail.com.

Background and Objective: Coeliac disease (CeD) is a common, underdiagnosed enteropathy with rising incidence and diagnostic delay. This literature review synthesises advances in small bowel capsule endoscopy (SBCE) and artificial intelligence (AI) for SBCE, and outlines implications for clinical practice.

Methods: A comprehensive literature search in PubMed, Scopus, Embase, and Cochrane Library was conducted, where relevant articles published in English over the past ten years [2015–2025] were selected and analysed by two independent reviewers.

Key Content and Findings: Current evidence supports tissue transglutaminase immunoglobulin A (tTG-IgA) as the first-line test and endomysial antibody IgA (EMA-IgA) as a test to rule in disease. An adult no-biopsy pathway at ≥10 times the upper limit of normal (ULN) yields near-perfect specificity but modest sensitivity; therefore, histology remains the reference standard. Optimised biopsy protocols with ≥4 samples from the second part of the duodenum plus 1–2 samples from the bulb, which are well-oriented, increase diagnostic yield. SBCE complements oesophagogastroduodenoscopy (OGD) to map disease extent, detect complications, and guide care when biopsy is contraindicated. A positive baseline study may be prognostic. AI has progressed from per-frame villous atrophy (VA) detection (internal accuracy: 94–96%) to patient-level and severity curve methods showing high agreement with experts, enabling reproducible burden mapping. Across prospective studies and meta-analyses in mixed SBCE indications, AI assistance increases sensitivity without losing specificity and reduces review time approximately 10–12-fold. Gains are greatest for non-experts and for triage applications. Key limitations include small, single-centre datasets, inconsistent labelling, image frame analysis rather than full videos, data leakage risks, and uncertain generalisability across devices and populations. Priorities include multicentre, patient-wise external validation; harmonised International Capsule Endoscopy Research (I-CARE) lesion definitions; prevalence-aware calibration; equity-aware evaluation; and vendor-agnostic deployment.

Conclusions: AI-augmented SBCE can improve efficiency, consistency, and monitoring of CeD; however,

^ ORCID: 0000-0002-1794-2569.

adoption should remain human-in-the-loop and be anchored to safety protocols, including patency testing when retention risk is relevant. Equity considerations include serology-negative presentations in some populations and the need for calibrated thresholds aligned with real-world prevalence and costs.

Keywords: Celiac disease; capsule endoscopy; artificial intelligence (AI); villous atrophy (VA); small intestine

Received: 13 September 2025; Accepted: 05 December 2025; Published online: 26 January 2026.
doi: 10.21037/tgh-25-128
View this article at: https://dx.doi.org/10.21037/tgh-25-128

Introduction

Background

Coeliac disease (CeD) is a chronic, immune-mediated enteropathy triggered by gluten in genetically susceptible individuals and has substantial implications for patient outcomes and healthcare systems. CeD is common worldwide, yet a substantial proportion of cases remain undiagnosed or are diagnosed after a prolonged delay, leading to ongoing symptoms and an impaired quality of life (1-4). Current diagnostic pathways rely on serological testing and histological confirmation from duodenal biopsies obtained at oesophagogastroduodenoscopy (OGD), which are invasive, resource-intensive, and vulnerable to sampling error and inter-observer variation (5,6).

Rationale and knowledge gap

Integrating artificial intelligence (AI) with small bowel capsule endoscopy (SBCE) offers promising solutions to longstanding diagnostic limitations whilst raising essential questions about implementation, safety, and equity. However, the evidence base for AI-assisted SBCE in CeD is still fragmented, with many studies being small, single-centre, and using heterogeneous outcome measures and reference standards. As a result, clinicians and researchers lack a concise synthesis of how AI-augmented SBCE performs in detecting and quantifying villous atrophy (VA), how it influences workflow and reading time, and what safety, equity, and generalisability issues need to be addressed before wider adoption.

Objective

This narrative review consolidates advancements in SBCE and AI for SBCE, discussing their implications for clinical practice. Specifically, we aim to summarise current diagnostic pathways for CeD, appraise the available literature on AI applications in SBCE with regard to accuracy, workflow, and safety, and identify key gaps and priorities for future research and clinical implementation. We present this article in accordance with the Narrative Review reporting checklist (available at https://tgh.amegroups.com/article/view/10.21037/tgh-25-128/rc).

Methods

To compile this narrative review, we conducted a comprehensive literature search in PubMed, Scopus, Embase, and Cochrane Library using various combinations and variations of the following keywords: “Coeliac disease”, “Artificial intelligence”, “Small bowel”, and “Capsule endoscopy”. The search was limited to articles published in English and included original studies investigating the integration of AI in CeD. Animal studies, studies focusing on diseases other than CeD, and articles not published in English were excluded. We selected and analysed articles published up to August 15, 2025, with a specific focus on recent articles published over the past decade [2015–2025]. Screening and selection of studies were done by two independent reviewers (R.M. and A.D.), followed by full-text analysis. The details of the search strategy for this review are provided in *Table 1*.

Epidemiology

Globally, the pooled serology-based prevalence of CeD is ~1.4%, and the biopsy-confirmed prevalence is ~0.7%, with higher rates in women (≈1.5 times) and in children (≈2 times) compared with adults (1). Proposed mechanisms include differences in immune function (higher $CD4^+$ and $CD8^+$ activity), X-linked gene variants, hormonal influences such as oestrogen and androgens, and life events such as pregnancy or menstruation that can alter intestinal permeability and immune regulation. These combined genetic, hormonal, and environmental factors may

Table 1 Summary of our search strategy

Items	Specification
Date of search	August 15, 2025
Databases and other sources searched	PubMed, Scopus, Embase, and Cochrane Library
Search terms used	"Coeliac Disease", "Celiac Disease", "Coeliac Sprue", "Celiac Sprue", "Gluten Enteropathy", "Nontropical Sprue", "Artificial Intelligence", "AI (Artificial Intelligence)", "Computational Intelligence", "Machine Intelligence", "Machine Learning", "Small Intestine", "Small Bowel", "Capsule Endoscopy", "Video Capsule Endoscopy", "Wireless Capsule Endoscopy"
Timeframe	Articles published up to August 15, 2025, with a specific focus on recent articles published over the past decade [2015–2025]
Inclusion and exclusion criteria	Inclusion criteria: (I) original studies investigating the integration of AI in coeliac disease; (II) articles published in English. Exclusion criteria: (I) animal studies; (II) studies focusing on diseases other than coeliac disease; (III) articles not published in English
Selection process	Screening and selection of studies were done by two independent reviewers (R.M. and A.D.). Any conflicts were resolved by a third reviewer (J.B.)

AI, artificial intelligence.

contribute to both higher disease susceptibility and greater intestinal injury in women with CeD (2). Underdiagnosis remains pervasive. The fourth population-based survey in the Trøndelag Health Study (HUNT4) in Norway revealed that 75% of individuals with CeD were newly identified through screening, and most reported symptomatic/quality-of-life gains after a gluten-free diet (3). "Hidden" prevalence has been highlighted in US data showing that Black patients may present with serology-negative yet biopsy-confirmed disease (4). Together, these findings support the contemporary concept of the "coeliac iceberg". This underdiagnosis justifies interest in non-invasive visual modalities, such as SBCE, where AI support may help expose overlooked diseases and improve equity in detection.

Current diagnostics

Serology: assay performance and adult "no-biopsy" thresholds

With respect to first-line tests, a contemporary meta-analysis in adults shows that tissue transglutaminase immunoglobulin A (tTG-IgA) has a sensitivity of 90.7% [95% confidence interval (CI): 87.3–93.2%] and specificity of 87.4% (84.4–90.0%), while endomysial antibody IgA (EMA-IgA) has a sensitivity of 88.0% (75.2–94.7%) and specificity of 99.6% (92.3–100%) (5). These data support tTG-IgA as a screening test and EMA-IgA as a means to "rule in" CeD. Total IgA should be measured to detect IgA deficiency; IgG-based assays are alternatives when IgA is low (6).

Adult no-biopsy approach

In a UK multicentre study, tTG-IgA levels ≥10 times the upper limit of normal (ULN) identified adults with histology diagnostic of CeD with very high positive predictive value (PPV), supporting a no-biopsy pathway in carefully selected patients (7). A 2024 meta-analysis of 18 studies encompassing 12,103 adults found a pooled specificity of 100% (95% CI: 98–100%) and a sensitivity of 51% for the ≥10 times the ULN threshold, with modelled PPV of 95–99% at pretest prevalences of 10–40% (8). Most guidelines still require biopsy in most adults; no-biopsy pathways should be limited to high pretest probability scenarios and robust assays, often with EMA confirmation (6).

Endoscopy and biopsy: sampling strategies that reduce error

Because VA can be patchy, evidence and guidelines support ≥4 biopsies from the second part of the duodenum (D2) plus 1–2 from the bulb (D1); correct orientation improves interpretability (9). A 2018 meta-analysis showed that adding a bulb biopsy increased diagnostic yield by ~5% (10). A 2024 meta-analysis focused on D1 confirms the incremental yield in adults (11). Biopsies and serology

must be performed on a gluten-containing diet; endoscopic visualisation alone is insufficient to diagnose CeD in adults. Current American College of Gastroenterology (ACG) guidance continues to regard histology as the reference standard in most adult cases (6). Recent meta-analyses have collectively reinforced the central roles of serology, biopsy protocols, and no-biopsy thresholds in current diagnostic pathways. The pooled performance metrics of these approaches, including duodenal bulb biopsy yield, are summarised in *Table 2*, providing context for the evolving diagnostic algorithms.

The utility of SBCE in the diagnosis of CeD

SBCE can depict the extent of mucosal injury and detect complications; it is most useful when OGD/biopsy is contraindicated, non-diagnostic, and suspected complications (e.g., ulcerative jejunoileitis, strictures, refractory CeD). While ESGE guidelines support the role of SBCE in identifying complications and in cases of equivocal disease, they are against its use for disease monitoring (12). In equivocal adult CeD, positive SBCE at diagnosis has been associated with worse outcomes, suggesting prognostic value (13). In addition, several studies have shown the benefit of repeat SBCE in refractory CeD, demonstrating an improvement in the extent of disease post-treatment (14-16). Multicentre series and reviews report high per-frame accuracy for VA on SBCE; however, heterogeneity in datasets and reference standards means that SBCE does not replace histology for initial adult diagnosis (17). Overall capsule retention is ~2% [higher in suspected/known inflammatory bowel disease (IBD)] (18). Patency testing halves retention in high-risk groups (19). These figures guide pre-procedure risk assessment when strictures are possible. The current diagnostic pathways for CeD, including the evolving role of SBCE and potential AI augmentation, have been summarised in *Figure 1*.

AI for SBCE in CeD

From architecture choices to the unit of evaluation

The modern literature spans two main goals: (i) detecting VA on single frames or short clips; and (ii) estimating patient-level disease burden along the small bowel. Zhou *et al.* pioneered an accurate patient-level readout using a deep learning model called a convolutional neural network (CNN)—a type of algorithm that automatically learns image features such as texture or pattern differences—to analyse SBCE clips after a pre-processing pipeline that removed blurred frames and normalised lighting (20). The authors used GoogleNet, a CNN developed by Google, for their analysis. They achieved perfect sensitivity and specificity in a small pilot study, showing that an AI model can capture disease features visible to expert readers. Methodologically, the study is notable for: (I) making rotation/illumination variance an explicit design target; and (II) reporting patient-level performance—albeit on a very small, single-centre cohort.

Subsequent work refined this approach by utilising more advanced "attention" modules that help the network focus on key areas of each frame. Wang *et al.* introduced a channel-based recalibration layer, Block-wise Channel Squeeze-and-Excitation (BCSE), within a standard CNN backbone to enhance feature discrimination (21). Rather than relying solely on one classifier, they tested various mathematical classifiers, such as support vector machines (SVMs), simpler k-nearest neighbour (KNN) and linear discriminant analysis (LDA) methods, which essentially group images based on learned patterns. The best setup reached an accuracy of 95.94%, a sensitivity of 97.20%, and a specificity of 95.63%. The authors explicitly acknowledge the risk of data leakage when adjacent frames from the same study are split across folds and call for patient-wise evaluation on larger cohorts. Technically, the gain came from channel-wise attention and decoupling the classifier head (SVM on deep features), which can improve calibration and robustness.

Demonstrating that heavy CNNs are not strictly necessary, Stoleru *et al.* designed a lighter, more interpretable system that utilised hand-engineered image features such as brightness, edge sharpness, and tissue texture to detect VA (22). They reported an accuracy of 94.1% on SBCE frames. While still per-frame and single-source, this result is significant for deployment, as it suggests that centres without Graphics Processing Unit (GPU) infrastructure can still approach deep-net performance using interpretable texture/edge descriptors. This approach achieved performance comparable to deep learning models while requiring less computing power, suggesting practical feasibility even in resource-limited settings.

Across methods, the per-frame discrimination of VA from normal mucosa is consistently high (~94–96% internal accuracy), but patient-level generalisation remains under-reported beyond small, single-centre tests. Architecture and

Table 2 Diagnostic pathway accuracy studies (serology or histology)

Study (first author, year)	Design/population	Key accuracy metrics	Outcomes
Sheppard *et al.*, 2022 (5)	Systematic review & meta-analysis (adults)—serologic tests *vs.* biopsy	tTG-IgA: Sens 90.7%, Spec 87.4%. EMA-IgA: Sens 88.0%, Spec 99.6%	Supported IgA tTG as first-line (high sensitivity) and EMA as confirmatory (very high specificity) for coeliac diagnosis. Total IgA level must be checked to exclude IgA deficiency
Penny *et al.*, 2021 (7)	Multicenter observational (UK + international cohorts)—adults with high tTG titres	IgA tTG ≥10× ULN had PPV ~95–100% for Marsh 3 histology across cohorts. Sensitivity ranged ~30–54% and specificity ~83–100% depending on cohort	Demonstrated that very high tTG-IgA levels (≥10× upper limit) strongly predict villous atrophy in adults (nearly perfect PPV), albeit with modest sensitivity. Provides evidence for a no-biopsy diagnosis in carefully selected adults
Shiha *et al.*, 2024 (8)	Systematic review & meta-analysis (18 studies, 12,103 adults)—no-biopsy threshold	Pooled Sens 51%, Spec 100% for tTG ≥10× ULN. Modelled PPV ~95–99% at pre-test probabilities 10–40%	Confirms that the ≥10× tTG-IgA threshold yields near-absolute specificity (≈100%) for histologic CeD in adults, but detects only ~50% of cases. Emphasises biopsy is still required in ~half of adults (especially in lower-titer or atypical cases)
McCarty *et al.*, 2018 (10)	Systematic review & meta-analysis—duodenal bulb biopsy yield	Adding a duodenal bulb biopsy resulted in a ~5% increase (95% CI: 3–9%) in diagnostic yield for CeD	Validates guidelines recommending ≥1–2 biopsies from the duodenal bulb (in addition to ≥4 from D2) to detect patchy villous atrophy. The incremental yield (~5%) can be clinically significant in borderline cases
Deb *et al.*, 2024 (11)	Systematic review & meta-analysis (adults)—role of duodenal bulb biopsy	Adding a D1 biopsy resulted in a ~6.9% increase (95% CI: 4.6–10.2%) in diagnostic yield for CeD	Reinforces that duodenal bulb biopsies enhance detection of CeD in adults, aligning with prior evidence. Bulb involvement may be the only site of lesion in a subset of patients, justifying routine bulb sampling

CeD, coeliac disease; CI, confidence interval; EMA, endomysial antibody; IgA, immunoglobulin A; PPV, positive predictive value; Sens, sensitivity; Spec, specificity; tTG, tissue transglutaminase antibody; ULN, upper limit of normal.

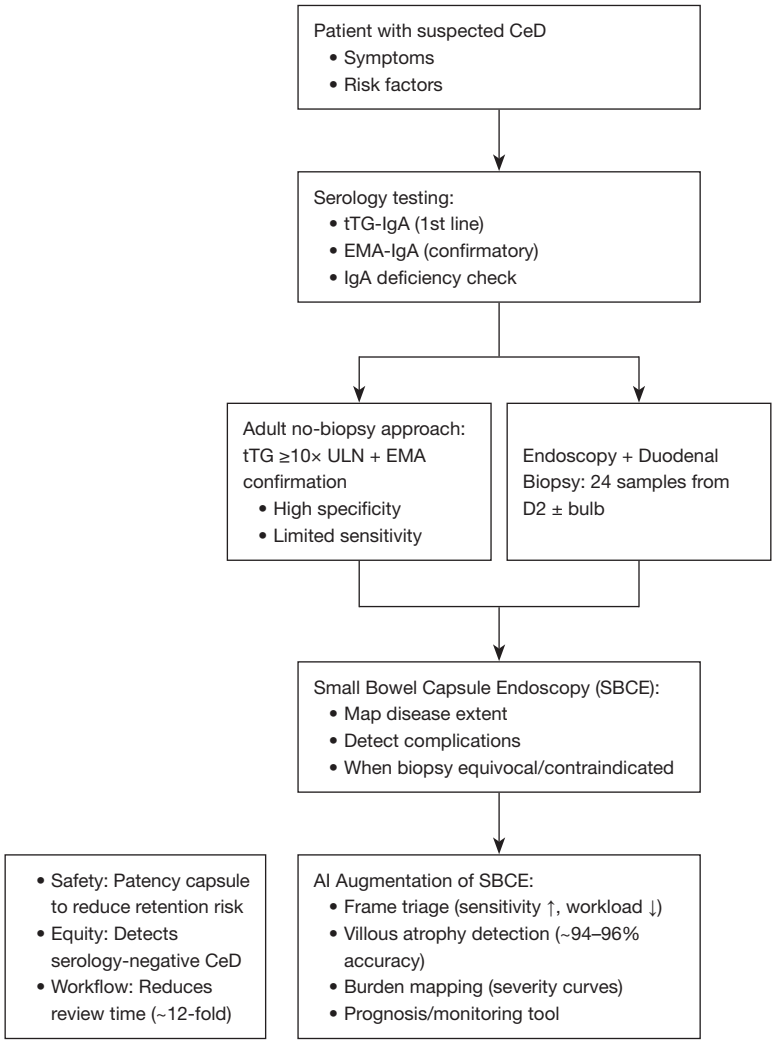

Figure 1 Diagnostic pathway flowchart with AI integration. AI, artificial intelligence; CeD, coeliac disease; D2, second part of duodenum; EMA, endomysial antibody; IgA, immunoglobulin A; SBCE, small-bowel capsule endoscopy; tTG, tissue transglutaminase antibody; ULN, upper limit of normal.

classifier choices can contribute measurable gains, yet their true value must be judged under patient-wise splits and external validation. Methodologically, CNNs and related models show that AI can recognise mucosal changes at the pixel level and summarise them consistently across patients, offering a reproducible framework for objective SBCE interpretation in CeD.

From detection to measurement: severity quantification and disease mapping

Clinically, SBCE is often used to stage disease extent and severity rather than merely flag the presence or absence of VA. Two linked studies address this quantitative agenda. In 2020, Chetcuti Zammit *et al.* used a probabilistic classifier trained on explicitly defined SBCE features aligned with recognised atrophic signs (scalloping, mosaic pattern, nodularity, fissuring, ulcers) to (i) predict Marsh severity and (ii) separate CeD from serology-negative VA (SNVA) (23). Internal validation yielded 69.1% accuracy for both tasks; incorporating class-proportion priors (prevalence-aware modelling) improved CeD *vs.* SNVA discrimination to 75.3%. These targets are deliberately more challenging than "CeD *vs.* normal" frame classification, reflecting real-

world differentials and the imperfect coupling between macroscopic atrophy and histology. In 2023, the same group advanced a continuous severity-curve approach (24). In 63 biopsy-proven adults, three expert readers scored frames on an ordinal 0–3 scale; a trained machine-learning algorithm (MLA) generated parallel scores, which were aggregated into segmental (tertiles) and whole-bowel curves. Inter-reader agreement on the whole-bowel mean severity was Krippendorff's α=0.924 (excellent). Crucially, reader *vs.* MLA agreement was similarly high: α=0.945 for the whole-bowel mean, and in the proximal tertile, α=0.932 (mean) and 0.867 (maximum). By transitioning from binary detection to standardised, reproducible quantification, this paradigm directly supports longitudinal monitoring, objective reporting (particularly in low-volume and non-specialist settings), and potentially therapy decisions in CeD. Limitations are the single-centre origin and the absence of an external replication set using the same scoring protocol. The performance characteristics of AI models applied specifically to CeD SBCE datasets, including both frame-level and patient-level approaches, are summarised in *Table 3*.

Human-algorithm comparison: *what does "expert-level" mean here?*

Direct, task-matched comparisons in CeD on SBCE are uncommon. The severity-curve study (24) provides the best head-to-head signal, i.e., MLA recapitulated expert severity curves with α up to 0.945 for whole-bowel mean scores and 0.867–0.932 proximally, indicating that an automated pipeline can produce clinically legible outputs (mean/max severity by segment) that mirror those of senior readers. Unlike frame-level accuracy, agreement statistics on ordinal scales capture what clinicians need from SBCE—consistent burden estimation—and should be prioritised in future evaluations.

External validation, *dataset diversity*, *and label standards*

A consistent limitation across this literature is the absence of rigorous external, multicentre validation with patient-level splits. Zhou *et al.* reported patient-level performance but evaluated only 10 test patients processed within the same pipeline, restricting generalisability (20). Wang *et al.* demonstrated excellent per-frame metrics under repeated 10-fold cross-validation; however, they provided no patient-wise or external evaluation (21). Likewise, Stoleru *et al.* (22) used single-source frames without an external test set, and the Chetcuti Zammit studies (13,24) relied on single-centre cohorts with internal validation only. Together, these designs risk optimistic estimates and do not establish robustness across centres, devices, or disease spectra. Systematic review of CeD AI echoes these issues (25)—small datasets, heterogeneous labels, and inconsistent reference standards—and summarises performance ranges (accuracy 84% to 95.94% in deep-learning studies; GoogLeNet's 100%/100% on a tiny patient-level test) while urging larger, more diverse cohorts and better detection of milder disease. One practical enabler is label harmonisation. The International Capsule Endoscopy Research (I-CARE) Delphi consensus standardised SBCE atrophic-lesion nomenclature (e.g., scalloping, nodularity/granularity, mosaic pattern, loss of folds), providing operational definitions that can underpin multicentre annotation and device-agnostic training (26). The adoption of CeD-AI pipelines is expected to reduce inter-annotator noise and enhance transportability.

Clinical translation: where AI for SBCE can help in suspected or confirmed CeD

Triage: *prioritising "high-yield" SBCEs and readers' attention*

AI triage fits naturally before a full human read, an auxiliary model pre-screens the stream, surfaces frames with atrophic features (e.g., scalloping, mosaic pattern), and suppresses long stretches of normal mucosa. In SBCE, more broadly, deep-learning auxiliary readers have shown large review-time reductions while maintaining or improving sensitivity. In a 2024 multicentre prospective study of bleeding lesions, AI-assisted reading was non-inferior and statistically superior to standard reading in terms of per-patient diagnostic yield (73.7% *vs.* 62.4%; P=0.0213), with reading time tracked as a secondary endpoint (27). A 2025 meta-analysis of proprietary AI add-ons reported pooled per-patient sensitivities with AI 0.93–1.00 *vs.* 0.75–0.89 for conventional reading, with no loss in specificity, and a 12-fold mean reduction in reading time (4.7 min AI-assisted *vs.* 56.7 min standard) (28). Images requiring review fell by a factor of 24 to 51 across the included studies. These findings support the use of AI as a front-end triage tool to focus expert time on ambiguous or obviously abnormal sequences. For CeD specifically, frame-level detectors achieve ~94–96% accuracy internally, and a small patient-

Table 3 Capsule endoscopy in coeliac disease, AI-based study summary

Study (first author, year)	Country; design	Population (n)	AI task (model)	Capsule device	Sensitivity	Specificity	Accuracy	Other outcomes
Zhou *et al.*, 2017 (20)	China & USA; retrospective	CeD patients *vs.* controls (11 *vs.* 10)	Frame-level villous atrophy detection (GoogLeNet CNN)	PillCam SB2	100%	100%	–	Achieved 100% sensitivity and specificity on test set (5 CeD, 5 controls); confidence metric correlated with disease severity
Wang *et al.*, 2020 (21)	China & USA; retrospective	CeD patients *vs.* healthy (12 *vs.* 13 adults; 2140 images)	Frame-level VA classification (ResNet50 + attention)	PillCam SB2 & SB3	97.2%	95.6%	95.9%	10×10-fold cross-validation results: high accuracy (≈96%) in classifying villous atrophy *vs.* normal
Stoleru *et al.*, 2022 (22)	Romania; retrospective	CeD patients *vs.* healthy (65 *vs.* 45; 109 films)	Patient-level CeD diagnosis (filter + SVM classifier)	PillCam SB3	–	–	94.1%	Achieved 94.1% accuracy (linear SVM) in test set; F1-score ~94%. AI used handcrafted features (mucosal atrophy patterns) rather than deep CNN
Chetcuti Zammit *et al.*, 2020 (23)	UK; retrospective	CeD patients with villous atrophy (incl. seronegative)	Patient-level severity prediction & CeD *vs.* SNVA differentiation (probabilistic ML model)	SBCE images	–	–	~69%	Higher Marsh score patients had more extensive SBCE lesions. Validation accuracy ~69.1% for Marsh severity and ~69.1% for CD *vs.* seronegative VA, rising to 75.3% when accounting for class prevalence
Chetcuti Zammit *et al.*, 2023 (24)	UK & USA; prospective comparison	CeD adults (63) with SBCE	Whole-small-intestine disease burden mapping (ordinal 0–3 severity scoring by AI)	SBCE images	–	–	–	Inter-observer agreement: Krippendorff's α=0.924 among 3 experts, and AI *vs.* experts $\alpha \approx 0.93$ (almost perfect). AI severity "curves" closely matched human readers, suggesting AI can reproducibly grade disease extent

AI, artificial intelligence; CeD, coeliac disease; CNN, convolutional neural network; ML, machine learning; SBCE, small-bowel capsule endoscopy; SNVA, serology-negative villous atrophy; SVM, support vector machine; VA, villous atrophy.

level study reported 100%/100% sensitivity/specificity, indicating that a triage filter can reliably elevate VA-like frames for expedited human review in suspected CeD. Beyond coeliac-specific work, AI has been widely evaluated in pan-indication SBCE studies, especially for triage and workflow optimisation. A summary of representative multicentre trials and meta-analyses is provided in *Table 4*.

Adjunct diagnosis: *complementing serology and histology in equivocal adults*

SBCE is already recommended as an adjunct when histology is contraindicated, non-diagnostic, or when disease extent/complications are suspected. AI can strengthen this role by providing consistent, reader-agnostic detection of VA patterns and highlighting segments for targeted histology when subsequent procedures are feasible. In head-to-head comparisons across SBCE indications, AI assistance enhances sensitivity without inflating false positives and particularly benefits non-expert readers. In a meta-analysis, AI-assisted juniors outperformed expert performance in conventional mode (sensitivity 99.2% *vs.* 91.1% in one included study) (28). In CeD-specific work, probabilistic models distinguishing CeD from SNVA reached 69.1% accuracy, improving to 75.3% after class-proportion priors were applied, illustrating the value of prevalence-aware calibration when moving from curated datasets to mixed-prevalence clinics. AI support may help most adults with high pre-test probability, but equivocal serology or histology, where AI-flagged VA segments, including the bulb, can justify repeat or directed biopsies when appropriate. In patients for whom biopsy is contraindicated, AI-augmented SBCE can increase diagnostic confidence, while acknowledging that histology remains the reference standard for most adults. Performance remains modest for differentiating celiac disease from seronegative VA; however, probabilistic scoring offers more explicable outputs that can be weighed alongside serology, HLA status, and clinical context.

Safety: *capsule retention risk and patency strategy*

The rare but consequential risk of capsule retention dominates SBCE safety. A 2017 meta-analysis reported overall retention of ≈2%, rising to ~4% in suspected and ~8% in known IBD; patency capsule or computed tomography (CT) enterography prior to SBCE halved retention risk in high-risk cohorts (18). Contemporary series emphasise patency testing as a pragmatic gatekeeper, with extended protocols improving availability at the expense of longer pre-procedure time (32). In contrast to Crohn's disease, small-bowel strictures are exceedingly uncommon in CeD, typically in refractory cases (33). For this reason, the baseline capsule retention risk is low; nonetheless, AI triage/monitoring does not alter mechanical risks, so standard safety pathways (patency imaging/testing when clinically indicated) remain unchanged. Multiple studies have quantified the baseline risk of capsule retention and the value of patency testing in high-risk populations. These data, including systematic reviews and recent extended-protocol studies, are outlined in *Table 5*, supporting risk-stratified use of SBCE in clinical practice.

Workflow and reading-time implications: *service load*, *training*, *and quality assurance (QA)*

SBCE reading is time-intensive and vulnerable to fatigue-related misses; AI integration occurs at two pressure points: frame selection and first-pass classification. In the 2025 meta-analysis of proprietary systems, AI assistance reduced review time from 56.7 to 4.7 min on average (a 12-fold increase) while increasing pooled sensitivity at unchanged specificity (28). The image burden presented to readers dropped by 24–51 times, allowing for concentration on ambiguous frames (28). A 2019 deep-learning model reported higher sensitivity with significantly shorter reading times than conventional analysis; subsequent multicentre work and randomized controlled trial (RCT)-style protocols have reinforced the efficiency gains (29). In real-world usage of a validated AI platform, centres documented substantial time savings with concordant detection performance *vs.* standard reading, supporting feasibility outside trial conditions (30). Together, these data justify positioning AI as a workload buffer in high-volume services and a training equaliser—in pooled analyses, AI-assisted junior readers matched or exceeded unassisted experts on sensitivity (28). The overall AI workflow for SBCE in CeD is illustrated in *Figure 2*, highlighting key stages from pre-processing to clinical output.

Equity and generalisability: *who gets detected*, *and where does AI work?*

The hidden burden is well documented. The HUNT4 population screening revealed that most CeD cases were previously undiagnosed, with symptomatic and quality-

Table 4 Pan-indication SBCE AI/workflow studies (mixed populations or triage)

Study (first author, year)	Indication/task	Sample size	AI System (model)	Sensitivity (AI *vs.* standard)	Reading time (AI *vs.* standard)
Spada *et al.*, 2024 (27)	Suspected small-bowel bleeding (multicentre prospective)	133 patients	NaviCam SB with "ProScan" CNN triage	73.7% *vs.* 62.4% detection yield for bleeding lesions (non-inferior and superior to manual)	3.8 *vs.* 33.7 min mean review time (~10× faster). AI missed fewer lesions (6.6% *vs.* 21% miss rate) and detected more lesions per patient
Cortegoso Valdivia *et al.*, 2025 (28)	AI *vs.* conventional SBCE reading (systematic review & meta-analysis)	6 studies (multi-device)	Various proprietary AI assist systems	Higher accuracy & sensitivity with AI across studies (pooled diagnostic OR 10.3 *vs.* 7.4). E.g., junior readers with AI achieved ~99% Sens *vs.* ~88–91% for experts without AI	~4.7 *vs.* 56.7 min pooled mean reading time (≈12-fold reduction) with AI assistance, demonstrating markedly improved efficiency
Ding *et al.*, 2019 (29)	Multi-disease SBCE detection (GI lesions classifier)	5,000 patients (validation)	Deep CNN (Ankon, 12-class model)	99.9% *vs.* 74.6% per-patient sensitivity (AI *vs.* doctors) for abnormalities; Spec ~100% *vs.* ~87%	5.9 *vs.* 96.6 min average reading time per case (AI *vs.* human). AI detected ~4,206 lesions with 99.9% sensitivity, far outperforming conventional review
O'Hara *et al.*, 2023 (30)	Real-world SBCE (OMOM capsule) in mixed GI cases (retrospective)	40 patients	OMOM AI software (CNN)	98.1% *vs.* 86.2% per-lesion detection of significant findings (AI *vs.* standard)—AI caught >98% of lesions *vs.* ~86% by manual reading	2.3 *vs.* 29.7 min mean reading time (AI *vs.* manual), saving ~27.4 minutes per study. Overall diagnostic conclusions were 100% concordant between AI-assisted and standard reads
Mascarenhas *et al.*, 2024 (31)	Multi-device vascular lesion detection (angioectasia, etc.)	1,022 exams	CNN trained on 7 capsule models (pan-endoscopic)	86.4% (AI model sensitivity); 98.3% specificity for vascular lesions in validation	N/A (algorithm processes ~115 frames/second)—not a live reader study. Overall accuracy 95.0%, PPV 95.2%, NPV 95.0%. Demonstrated cross-device interoperability of the AI

AI, artificial intelligence; CNN, convolutional neural network; GI, gastrointestinal; N/A, not applicable; NPV, negative predictive value; OMOM, OMOM HD Capsule Endoscopy system [Jinshan Science & Technology (Group), Yubei, China]; OR, odds ratio; PPV, positive predictive value; SB, small bowel; SBCE, small-bowel capsule endoscopy; Sens, sensitivity; Spec, specificity.

Table 5 Safety studies (capsule retention and patency)

Study (first author, year)	Context	Sample	Retention rate/outcome	Findings
Rezapour *et al.*, 2017 (18)	Systematic review & meta-analysis—capsule retention risk	25 studies (approx. 5,000 patients)	~1.4–2% overall capsule retention rate (all indications)	Overall retention ~2% in SBCE. Higher risk in known Crohn's or suspected strictures (up to ~4–5% in some subgroups). Emphasises risk stratification before SBCE (e.g., use patency capsule in high-risk patients)
O'Hara *et al.*, 2023 (19)	Retrospective (Ireland)—Patency capsule use *vs.* retention	152 patency tests	0% retained capsules among those cleared by patency test (*vs.* several retentions historically without testing)	Patency capsule screening prevented capsule retention in high-risk cases. However, ~15–20% of patients had patency capsule "fails" (false positives), leading to procedure deferral. Authors note that patency testing improves safety but can incur delays/cost (the "at what cost?" issue)
O'Hara *et al.*, 2024 (32)	Prospective study (Ireland)—72-hour patency capsule protocol	135 high-risk patients	57.8% confirmed functional patency (*vs.* 48.9% under 28 h standard); 0% retention in those who passed patency	Extending the patency monitoring window to 72 h significantly increased the fraction of high-risk patients cleared for SBCE (from ~49% to ~58%). No capsule retentions occurred in patients who passed the extended patency test, and roughly half of those who underwent CE had clinically significant findings. The 72 h protocol was safe and cost-neutral, avoiding unnecessary exclusions due to false-positive 28 h patency failures

CE, capsule endoscopy; SBCE, small-bowel capsule endoscopy.

Figure 2 Coeliac disease diagnostic and monitoring pyramid. AI, artificial intelligence; SBCE, small-bowel capsule endoscopy.

of-life gains on a gluten-free diet after case-finding (3). Diagnostic equity is a specific concern: in a US tertiary centre, Black patients with histology-consistent CeD were more likely to have negative serology, potentially excluding them from current adult pathways; an accompanying editorial highlighted algorithmic and systems contributors to missed diagnoses (4,34). In this context, a visual, morphology-based adjunct like SBCE—especially when AI-assisted—may help surface disease in groups at risk of serology-negative presentations, provided thresholds and outputs are calibrated to realistic prevalences.

Generalizability across devices and centres. SBCE AI research historically used single-centre, single-vendor datasets. Newer work demonstrates multi-brand, multi-device training for lesion detection across pan-endoscopic capsules, addressing technological interoperability—a prerequisite for wide rollout (31). Equally, label standardisation matters: the I-CARE Delphi consensus provides operational definitions for atrophic SBCE lesions (scalloping, mosaic pattern, loss of folds, nodularity). Prospective adoption should reduce inter-annotator noise and support cross-site learning and testing (26).

Limitations and future directions

Current AI applications in SBCE for CeD face several significant limitations that must be addressed for successful clinical implementation. These include small, single-centre datasets that limit generalisability, inconsistent labelling protocols across studies, risks of data leakage in model development, and uncertain performance across different capsule devices and patient populations.

Future research priorities should focus on multicentre, patient-wise external validation studies, developing harmonised I-CARE lesion definitions, implementing prevalence-aware calibration methods, equity-aware evaluation frameworks, and vendor-agnostic deployment strategies. Spatial mapping of mucosal changes in CeD using colour-coded small bowel maps to illustrate the distribution and severity of VA is a promising future direction. Similar frameworks have been applied in ulcerative colitis to depict segmental disease burden (35,36). Development of large-scale, publicly available annotated datasets in CeD is essential to enable such advances.

Conclusions

AI-augmented SBCE represents a promising advancement in CeD diagnosis and monitoring, offering potential improvements in efficiency, consistency, and accessibility. However, successful implementation requires careful attention to safety protocols, equity considerations, and the maintenance of human oversight in clinical decision-making. As this technology evolves, continuous evaluation of its real-world performance and impact on patient outcomes will be essential.

Acknowledgments

None.

Footnote

Reporting Checklist: The authors have completed the Narrative Review reporting checklist. Available at https://tgh.amegroups.com/article/view/10.21037/tgh-25-128/rc

Peer Review File: Available at https://tgh.amegroups.com/article/view/10.21037/tgh-25-128/prf

Funding: None.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tgh.amegroups.com/article/view/10.21037/tgh-25-128/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Singh P, Arora A, Strand TA, et al. Global Prevalence of Celiac Disease: Systematic Review and Meta-analysis. Clin Gastroenterol Hepatol 2018;16:823-836.e2.
2. Jansson-Knodell CL, Hujoel IA, West CP, et al. Sex Difference in Celiac Disease in Undiagnosed Populations: A Systematic Review and Meta-analysis. Clin Gastroenterol Hepatol 2019;17:1954-1968.e13.
3. Kvamme JM, Sørbye S, Florholmen J, et al. Population-based screening for celiac disease reveals that the majority of patients are undiagnosed and improve on a gluten-free diet. Sci Rep 2022;12:12647.
4. Cartee AK, Beasley TM, Estes D, et al. Black Patients Are More Likely to Have Negative Serologies With Celiac Consistent Histology in a Southeast US Tertiary Center. Gastro Hep Adv 2024;3:4-6.
5. Sheppard AL, Elwenspoek MMC, Scott LJ, et al. Systematic review with meta-analysis: the accuracy of serological tests to support the diagnosis of coeliac disease. Aliment Pharmacol Ther 2022;55:514-27.
6. Rubio-Tapia A, Hill ID, Semrad C, et al. American College of Gastroenterology Guidelines Update: Diagnosis and Management of Celiac Disease. Am J Gastroenterol 2023;118:59-76.
7. Penny HA, Raju SA, Lau MS, et al. Accuracy of a no-biopsy approach for the diagnosis of coeliac disease across different adult cohorts. Gut 2021;70:876-83.
8. Shiha MG, Nandi N, Raju SA, et al. Accuracy of the No-Biopsy Approach for the Diagnosis of Celiac Disease in Adults: A Systematic Review and Meta-Analysis. Gastroenterology 2024;166:620-30.
9. Raiteri A, Granito A, Giamperoli A, et al. Current guidelines for the management of celiac disease: A systematic review with comparative analysis. World J Gastroenterol 2022;28:154-75.

10. McCarty TR, O'Brien CR, Gremida A, et al. Efficacy of duodenal bulb biopsy for diagnosis of celiac disease: a systematic review and meta-analysis. Endosc Int Open 2018;6:E1369-78.
11. Deb A, Moond V, Thongtan T, et al. Role of Duodenal Bulb Biopsy in Diagnosing Suspected Celiac Disease in Adult Patients: A Systematic Review and Meta-analysis. J Clin Gastroenterol 2024;58:588-95.
12. Pennazio M, Rondonotti E, Despott EJ, et al. Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European Society of Gastrointestinal Endoscopy (ESGE) Guideline - Update 2022. Endoscopy 2023;55:58-95.
13. Chetcuti Zammit S, Schiepatti A, Aziz I, et al. Use of small-bowel capsule endoscopy in cases of equivocal celiac disease. Gastrointest Endosc 2020;91:1312-1321.e2.
14. Chetcuti Zammit S, Sanders DS, Cross SS, et al. Capsule endoscopy in the management of refractory coeliac disease. J Gastrointestin Liver Dis 2019;28:15-22.
15. Zammit SC, Elli L, Scaramella L, et al. Small bowel capsule endoscopy in refractory celiac disease: a luxury or a necessity? Ann Gastroenterol 2021;34:188-95.
16. Ferretti F, Branchi F, Orlando S, et al. Effectiveness of Capsule Endoscopy and Double-Balloon Enteroscopy in Suspected Complicated Celiac Disease. Clin Gastroenterol Hepatol 2022;20:941-949.e3.
17. Luján-Sanchis M, Pérez-Cuadrado-Robles E, García-Lledó J, et al. Role of capsule endoscopy in suspected celiac disease: A European multi-centre study. World J Gastroenterol 2017;23:703-11.
18. Rezapour M, Amadi C, Gerson LB. Retention associated with video capsule endoscopy: systematic review and meta-analysis. Gastrointest Endosc 2017;85:1157-1168.e2.
19. O'Hara F, Walker C, McNamara D. Patency testing improves capsule retention rates but at what cost? A retrospective look at patency testing. Front Med (Lausanne) 2023;10:1046155.
20. Zhou T, Han G, Li BN, et al. Quantitative analysis of patients with celiac disease by video capsule endoscopy: A deep learning method. Comput Biol Med 2017;85:1-6.
21. Wang X, Qian H, Ciaccio EJ, et al. Celiac disease diagnosis from videocapsule endoscopy images with residual learning and deep feature extraction. Comput Methods Programs Biomed 2020;187:105236.
22. Stoleru CA, Dulf EH, Ciobanu L. Automated detection of celiac disease using Machine Learning Algorithms. Sci Rep 2022;12:4071.
23. Chetcuti Zammit S, Bull LA, Sanders DS, et al. Towards the Probabilistic Analysis of Small Bowel Capsule Endoscopy Features to Predict Severity of Duodenal Histology in Patients with Villous Atrophy. J Med Syst 2020;44:195.
24. Chetcuti Zammit S, McAlindon ME, Greenblatt E, et al. Quantification of Celiac Disease Severity Using Video Capsule Endoscopy: A Comparison of Human Experts and Machine Learning Algorithms. Curr Med Imaging 2023;19:1455-662.
25. Sharif K, David P, Omar M, et al. Deep Learning in Coeliac Disease: A Systematic Review on Novel Diagnostic Approaches to Disease Diagnosis. J Clin Med 2023;12:7386.
26. Elli L, Marinoni B, Sidhu R, et al. Nomenclature and Definition of Atrophic Lesions in Small Bowel Capsule Endoscopy: A Delphi Consensus Statement of the International CApsule endoscopy REsearch (I-CARE) Group. Diagnostics (Basel) 2022;12:1704.
27. Spada C, Piccirelli S, Hassan C, et al. AI-assisted capsule endoscopy reading in suspected small bowel bleeding: a multicentre prospective study. Lancet Digit Health 2024;6:e345-53.
28. Cortegoso Valdivia P, Fantasia S, Kayali S, et al. Conventional small-bowel capsule endoscopy reading vs proprietary artificial intelligence auxiliary systems: Systematic review and meta-analysis. Endosc Int Open 2025;13:a25442863.
29. Ding Z, Shi H, Zhang H, et al. Gastroenterologist-Level Identification of Small-Bowel Diseases and Normal Variants by Capsule Endoscopy Using a Deep-Learning Model. Gastroenterology 2019;157:1044-1054.e5.
30. O'Hara FJ, Mc Namara D. Capsule endoscopy with artificial intelligence-assisted technology: Real-world usage of a validated AI model for capsule image review. Endosc Int Open 2023;11:E970-5.
31. Mascarenhas M, Martins M, Afonso J, et al. Deep learning and capsule endoscopy: Automatic multi-brand and multi-device panendoscopic detection of vascular lesions. Endosc Int Open 2024;12:E570-8.
32. O'Hara FJ, Costigan C, McNamara D. Extended 72-hour patency capsule protocol improves functional patency rates in high-risk patients undergoing capsule endoscopy. World J Gastrointest Endosc 2024;16:661-7.
33. Scarmozzino F, Pizzi M, Pelizzaro F, et al. Refractory celiac disease and its mimickers: a review on pathogenesis, clinical-pathological features and therapeutic challenges. Front Oncol 2023;13:1273305.
34. Hujoel I. Current Diagnostic Algorithms May Fail to

Identify Black Americans With Celiac Disease. Gastro Hep Adv 2024;3:134-5.
35. Stidham RW, Cai L, Cheng S, et al. Using Computer Vision to Improve Endoscopic Disease Quantification in Therapeutic Clinical Trials of Ulcerative Colitis. Gastroenterology 2024;166:155-167.e2.
36. Fan Y, Mu R, Xu H, et al. Novel deep learning-based computer-aided diagnosis system for predicting inflammatory activity in ulcerative colitis. Gastrointest Endosc 2023;97:335-46.

doi: 10.21037/tgh-25-128
Cite this article as: Dhali A, Maity R, Biswas J, Hann A, Sidhu R, Sanders DS. Artificial intelligence-augmented small bowel capsule endoscopy for coeliac disease: a literature review on accuracy, workflow, and safety. Transl Gastroenterol Hepatol 2026;11:27.