AI-assisted capsule endoscopy reading in suspected small bowel bleeding: a multicentre prospective study

Cristiano Spada*, Stefania Piccirelli*, Cesare Hassan, Clarissa Ferrari, Ervin Toth, Begoña González-Suárez, Martin Keuchel, Marc McAlindon, Ádám Finta, András Rosztóczy, Xavier Dray, Daniele Salvi, Maria Elena Riccioni, Robert Benamouzig, Amit Chattree, Adam Humphries, Jean-Christophe Saurin, Edward J Despott, Alberto Murino, Gabriele Wurm Johansson, Antonio Giordano, Peter Baltes, Reena Sidhu, Milan Szalai, Krisztina Helle, Artur Nemeth, Tanja Nowak, Rong Lin, Guido Costamagna

Summary

Background Capsule endoscopy reading is time consuming, and readers are required to maintain attention so as not to miss significant findings. Deep convolutional neural networks can recognise relevant findings, possibly exceeding human performances and reducing the reading time of capsule endoscopy. Our primary aim was to assess the non-inferiority of artificial intelligence (AI)-assisted reading versus standard reading for potentially small bowel bleeding lesions (high P2, moderate P1; Saurin classification) at per-patient analysis. The mean reading time in both reading modalities was evaluated among the secondary endpoints.

Methods Patients aged 18 years or older with suspected small bowel bleeding (with anaemia with or without melena or haematochezia, and negative bidirectional endoscopy) were prospectively enrolled at 14 European centres. Patients underwent small bowel capsule endoscopy with the Navicam SB system (Ankon, China), which is provided with a deep neural network-based AI system (ProScan) for automatic detection of lesions. Initial reading was performed in standard reading mode. Second blinded reading was performed with AI assistance (the AI operated a first-automated reading, and only AI-selected images were assessed by human readers). The primary endpoint was to assess the non-inferiority of AI-assisted reading versus standard reading in the detection (diagnostic yield) of potentially small bowel bleeding P1 and P2 lesions in a per-patient analysis. This study is registered with ClinicalTrials.gov, NCT04821349.

Findings From Feb 17, 2021 to Dec 29, 2021, 137 patients were prospectively enrolled. 133 patients were included in the final analysis (73 [55%] female, mean age 66·5 years [SD 14·4]; 112 [84%] completed capsule endoscopy). At per-patient analysis, the diagnostic yield of P1 and P2 lesions in AI-assisted reading (98 [73·7%] of 133 lesions) was non-inferior (p<0·0001) and superior (p=0·0213) to standard reading (82 [62·4%] of 133; 95% CI 3·6–19·0). Mean small bowel reading time was 33·7 min (SD 22·9) in standard reading and 3·8 min (3·3) in AI-assisted reading (p<0·0001).

Interpretation AI-assisted reading might provide more accurate and faster detection of clinically relevant small bowel bleeding lesions than standard reading.

Funding ANKON Technologies, China and AnX Robotica, USA provided the NaviCam SB system.

Lancet Digit Health 2024; 6: e345–53

*Joint first authors

Department of Medicine, Gastroenterology and Endoscopy (Prof C Spada MD PhD, S Piccirelli MD, D Salvi MD, Prof G Costamagna MD PhD) **and Unit of Research and Clinical Trials** (C Ferrari PhD), **Fondazione Poliambulanza Istituto Ospedaliero, Brescia, Italy; Università Cattolica del Sacro Cuore, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy** (Prof C Spada, S Piccirelli, D Salvi, Prof G Costamagna)**; IRCCS Humanitas Research Hospital, Department of Biomedical Sciences, Rozzano, Milan, Italy** (Prof C Hassan MD PhD)**; Skåne University Hospital, Lund University, Department of Gastroenterology, Malmö, Sweden** (E Toth MD PhD, G Wurm Johansson, A Nemeth)**; Hospital Clínic of Barcelona, Endoscopy Unit, Gastroenterology Department, Barcelona, Spain** (B González-Suárez MD PhD, A Giordano MD)**; Agaplesion Bethesda Krankenhaus Bergedorf, Academic Teaching Hospital of the University of Hamburg, Clinic for Internal Medicine, Hamburg, Germany** (M Keuchel MD PhD, P Baltes MD)**; Sheffield Teaching Hospitals NHS Trust, Academic Department of Gastroenterology and Hepatology, Sheffield, UK** (M McAlindon MD PhD, R Sidhu MD PhD)**; Department of Infection, Immunity and Cardiovascular Disease, University of Sheffield, Sheffield, UK** (M McAlindon, R Sidhu)**; Endo-Kapszula Health Centre and Endoscopy Unit, Department of Gastroenterology, Székesfehérvár, Hungary** (Á Finta MD, M Szalai MD)**; University of Szeged, Department of Internal**

Introduction

Small bowel capsule endoscopy is mainly indicated in patients with suspected small bowel bleeding, irrespective of sex.[1] In this context, small bowel capsule endoscopy shows about a 60% diagnostic yield, with angiodysplasias being the most common finding, accounting for 50% of such diagnoses.[2,3] Despite its high clinical feasibility and minimal invasive feature, the evaluation of capsule endoscopy video images is time consuming and requires the reader's ongoing concentration. Artificial intelligence (AI) has penetrated different fields in medicine, including gastrointestinal endoscopy.[4,5] Preliminary reports suggest that AI, in particular deep convolutional neural networks, are able to efficiently recognise specific images among a large variety, exceeding human performance in visual tasks.[6–18] Recently, a deep learning model (ProScan, Ankon, China and AnX Robotica, USA) was presented and validated in a retrospective series,[19] showing 99·88% sensitivity (per-patient analysis) and 99·90% sensitivity (per-lesion analysis) for the detection of small bowel abnormalities. AI-assisted reading also shortened the reading time (5·9 min [SD 2·2] *vs* 96·6 min [22·5], p<0·001).[19]

To date, no prospective multicentre trial has been published to confirm these preliminary results. The primary aim of this trial was to evaluate the non-inferiority of capsule endoscopy reading using ProScan versus standard reading in detection of significant small bowel pathology in patients with suspected small bowel bleeding. Subsequently, the superiority of capsule endoscopy reading using ProScan versus standard reading and the assessment of its impact on reading time were evaluated.

Medicine, Szeged, Hungary (A Rosztóczy MD PhD, K Helle MD); Sorbonne University, Saint Antoine Hospital, APHP, Centre for Digestive Endoscopy, Paris, France (Prof X Dray MD PhD); Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Digestive Endoscopy Unit, Rome, Italy (M E Riccioni MD PhD); Hôpital Avicenne, Université Paris 13, Service de Gastroenterologie, Bobigny, France (R Benamouzig MD PhD); South Tyneside and Sunderland NHS Foundation Trust, Gastroenterology, Stockton-on-Tees, UK (A Chattree MD); St Mark's Hospital and Academic Institute, Department of Gastroenterology, Middlesex, UK (A Humphries MD PhD); Hospices Civils de Lyon-Centre Hospitalier Universitaire, Gastroenterology Department, Lyon, France (Prof J-C Saurin MD PhD); The Royal Free Hospital and University College London (UCL) Institute for Liver and Digestive Health, Royal Free Unit for Endoscopy, London, UK (E J Despott MD PhD, A Murino MD); Medical Affairs, Hamburg, Germany (T Nowak PhD); Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Department of Gastroenterology, Wuhan, China (Prof R Lin MD PhD)

Correspondence to: Dr Stefania Piccirelli, Department of Medicine, Gastroenterology and Endoscopy, Fondazione Poliambulanza Istituto Ospedaliero, Brescia, Italy stefania.piccirelli@gmail.com

See Online for appendix

Research in context

Evidence before this study
Small bowel capsule endoscopy is mainly indicated in patients with suspected small bowel bleeding. Despite its high clinical feasibility and minimal invasive feature, the evaluation of capsule endoscopy videos is time consuming and requires reader's ongoing concentration. To help clinicians in this evaluation, artificial intelligence (AI) by means of deep convolutional neural networks appears to be a promising tool. However, use of deep convolutional neural networks in capsule endoscopy only started recently, and the available data and studies assessing the contribution of AI in improving the detection of small bowel abnormalities and in reducing the reading time are limited to retrospective study design. We searched PubMed between Jan 1, 2019, and Sept 30, 2023, for original research or review articles published in English using the search terms ('deep convolutional neural networks' AND 'capsule endoscopy' AND 'small bowel'). The search returned 28 articles: 21 articles concerned the validation of deep learning software and algorithms for lesion detections; three were reviews or opinion papers on potential application of AI in the gastroendoscopy field. Only four studies evaluated the use of AI in a clinical setting. Among these, just one was a multicentre (retrospective) study involving only 20 patients.

Added value of this study
To our knowledge, this is the first prospective, multicentre trial aiming to confirm preliminary results investigating the contribution of AI in identifying abnormalities of the small bowel in patients with suspected small bowel bleeding.

Our results provide novel and specific evidence of both the non-inferiority and superiority of capsule endoscopy AI-assisted reading versus standard reading in detection of specific types of lesions in patients with capsule endoscopy. In addition, our results confirm the impact of the use of AI in the remarkable reduction of the reading time.

Implications of all the available evidence
Although scarce data and pre-existing evidence are currently available, AI seems to play a role in improving diagnostic performances in capsule endoscopy. This prospective study provides first evidence of the efficiency of AI-assisted capsule endoscopy reading in reducing the reading time. Moreover, AI-assisted reading demonstrated efficacy and reliability, in terms of non-inferiority and superiority comparisons with standard reading, and in the detection of potentially bleeding lesions. Our results support and encourage the spread of AI features which can support clinicians during capsule endoscopy reading, ensuring a reliable diagnostic performance alongside a remarkable reduction of reading time, with a mitigation of the clinician burden.

Methods

Study design
This was a multicentre, prospective trial. A consecutive series of patients with suspected small bowel bleeding were enrolled at 14 European referral centres (appendix p 6). Inpatients and outpatients of male and female sex were included. Suspected small bowel bleeding was defined as anaemia with or without the presence of melena or haematochezia, and negative bidirectional endoscopy. The inclusion criteria for patients were as follows: aged 18 years or older, suspected small bowel bleeding, anaemia, defined with a haemoglobin cut-off of less than 13 g/dL for males and less than 11 g/dL for females, and a negative pregnancy test, when requested. Exclusion criteria are summarised in the appendix (p 7).

Capsule endoscopy protocol
Capsule endoscopy was performed using NaviCam SB system (Ankon, China and AnX Robotica, USA). The NaviCam SB system is provided with the Navicam SB capsule, the recording system (NS-I), and the workstation equipped with ESView (version 2.0) software and ProScan, a deep convolutional neural network-based feature. ProScan was trained and validated using 113 426 569 images from 6970 patients enrolled in a multicentre retrospective study by Ding and colleagues, involving 77 centres over a 2-year time period. In detail, ProScan was trained using 158 235 small bowel capsule endoscopy images from a subset of 1970 patients. The model was then validated using the remaining 5000 patients (no overlap of patients between training and validation). In the assessment of the small bowel, ProScan is able to perform a preliminary automated reading of capsule endoscopy videos, distinguishing abnormal findings (inflammation, ulcer, polyp, lymphangiectasia, bleeding, vascular disease, protruding lesion, lymphatic follicular hyperplasia, diverticulum, and parasite) from normal mucosa.[19] When using the ESView software for capsule endoscopy reading, the ProScan function can be activated or deactivated by clicking the correspondent button. Once activated, a shortened capsule endoscopy video is visualised, made up exclusively of frames containing the selected abnormalities which are marked with a blue box. In this way, the ProScan serves as a first automated reader, hiding all the irrelevant frames that would not be visualised by the user. For this reason, when performed, AI-assisted reading requires a very low speed (2–5 frames per s) to notice any pathological finding, and the remaining reading functions (ie, capture of a frame, pause, backward, and forward) remain the same as standard reading. Small bowel capsule endoscopy

examination was performed according to local rules. A split dose of very low or low dose PEG-based laxative was recommended as standard regimen. 4 h after capsule ingestion a light meal was allowed. Capsule endoscopy examination was completed if the capsule reached the colon within the recording time. Follow-up of patients was performed according to clinical routine, with collection of eventual cases of retention or adverse events.

Procedures

Assessment of videos consisted of consecutive steps and involved a total of 22 expert capsule endoscopy readers (at least 500 capsules read). First, at the site of patient's enrolment, investigators performed small bowel capsule endoscopy and evaluated the resulting video in standard mode according to the recommendations of the European Society of Gastrointestinal Endoscopy, at 10 frames per s in single-view mode in the small bowel, and 20 frames per s in the oesophagus or stomach, and colon.[20] Landmarks (first image of the gastrointestinal tract, first duodenal image, first caecal image) were manually selected by the reader; findings were captured by mouse click with the stopwatch running. After reading completion, findings were labelled specifying their bleeding potential (P0, P1, or P2, according to the adapted Saurin classification; appendix p 8) and their time of appearance (h: min: s, single frame or interval timing, as needed).[21] Location of findings was reported according to a simplified location protocol which divides the small bowel in three tertiles according to the small bowel transit time. Cleansing level was assessed as adequate or non-adequate according to the qualitative Brotz scale.[22]

Thereafter, investigators at the enrolment centres converted and anonymised the videos which were randomly reallocated to another centre for the second blinded AI-assisted reading (ie, the readers who performed the capsule endoscopy reading in the AI modality were unaware of the results of the first reading in the standard modality). The same instructions as specified previously were valid for the AI-assisted reading except for the reading speed, which was limited to 2–5 frames per s. The first reading was performed by the investigator of the centre where the patient was enrolled (not randomised). This was decided to guarantee the patients the standard of care of the centre, and free access to a complete clinical history and medical records for the physician who eventually diagnosed pathology. The second reading was performed by another reader from an external centre (randomly assigned). Finally, a board of expert (defined as having read more than 500 capsules) readers (CS, SP, EJD, MK, and BG-S) reviewed all videos to compare the results, and to evaluate the match of the findings reported by standard and AI-assisted readers. To minimise any interpretation bias of board members, no information regarding the original centre or the original reader was provided. The match of lesions was assessed for type (as defined in appendix p 8), and for the timing of the frame (h: min: s) using the original or the AI-converted video, accordingly. All P1 and P2 lesions were included in the analysis. P0 lesions were not considered since they do not impact clinical outcome. At per-patient analysis, standard and AI-assisted readings were considered concordant if the readings reported the same P2 lesion, irrespective of the P1 lesions (or the same P1 lesions in case no P2 lesions were reported). In case of multiple findings, the concordance was positive if at least one of the main findings was coincident. At per-lesion analysis, lesions were considered matching when they were described at the same timing or in 5 min intervals by the readers in each modality (standard and AI-assisted modes).

In case of discrepancies between initial readings (either standard or AI-assisted) and board readings, the board performed a consensus reassessment (adjudication). Discrepancies of findings reported by initial readers (either standard or AI-assisted) were observed in terms of presence or absence of lesions (at per-patient analysis) and in terms of the number of lesions (at per-lesion analysis). Discrepancies in terms of category of lesions were not observed (ie, lesions were categorised according to the Saurin classification modified for the protocol). AI performance was evaluated in a multistep approach at per-patient and per-lesion analyses. The board reading was used as gold standard to compare readings (AI-assisted and standard), and to measure accuracy. Standard reading was used as the reference to measure the accuracy of AI-assisted reading. Finally, the sensitivity of the ProScan system was measured by comparing standard and AI-assisted results, using the consensus reassessment (adjudication) as reference. Only investigators had access to the patient's clinical record, and the Good Data Protection Practice in Research guidelines were respected. Before capsule endoscopy procedures, each patient was asked to provide written informed consent.

Outcomes

The primary endpoint was to assess the non-inferiority of AI-assisted reading versus standard reading in the detection (diagnostic yield) of potentially small bowel bleeding P1 and P2 lesions in a per-patient analysis.

Secondary endpoints were as follows: non-inferiority of AI-assisted reading versus standard reading in detection of P1 and P2 lesions at per-lesion analysis; non-inferiority of AI-assisted reading versus standard reading in detection of P2 lesions at per-patient and per-lesion analysis; superiority of AI-assisted reading versus standard reading; accuracy of readers who used or did not use ProScan; sensitivity of the ProScan first automated reading; and mean reading time of AI-assisted versus standard reading.

Statistical analysis

Descriptive statistics by mean (SD) were computed for continuous variables, and frequency and percentage were

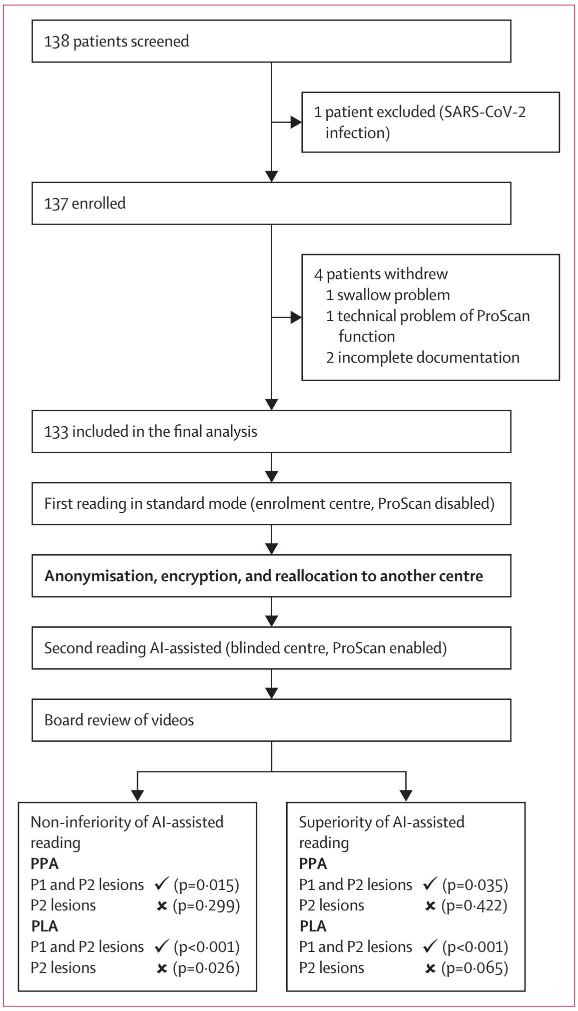

138 patients screened
1 patient excluded (SARS-CoV-2 infection)
137 enrolled
4 patients withdrew
1 swallow problem
1 technical problem of ProScan function
2 incomplete documentation
133 included in the final analysis
First reading in standard mode (enrolment centre, ProScan disabled)
Anonymisation, encryption, and reallocation to another centre
Second reading AI-assisted (blinded centre, ProScan enabled)
Board review of videos
Non-inferiority of AI-assisted reading
PPA
P1 and P2 lesions ✓ (p=0·015)
P2 lesions ✗ (p=0·299)
PLA
P1 and P2 lesions ✓ (p<0·001)
P2 lesions ✗ (p=0·026)
Superiority of AI-assisted reading
PPA
P1 and P2 lesions ✓ (p=0·035)
P2 lesions ✗ (p=0·422)
PLA
P1 and P2 lesions ✓ (p<0·001)
P2 lesions ✗ (p=0·065)

Figure 1: **Study design**
p value calculated at 95% CI. ✓=endpoint achieved. ✗=endpoint not achieved. AI=artificial intelligence. PPA=per-patient analysis. PLA=per-lesion analysis.

For more on **DTComPair** see https://cran.r-project.org/web/packages/DTComPair/DTComPair.pdf

provided for categorical variables. In detail, the accuracy performances of categorical binary variables were evaluated by assessing diagnostic yield, accuracy, sensitivity, specificity, positive predictive value and negative predictive value (a definition of these metrics is provided in the appendix p 3). Separate analyses, in per-patient and in per-lesion settings were performed for the evaluation of diagnostic accuracy and the other metrics in the two types of reading. In the per-patient analysis, paired sample tests (McNemar tests for proportion) and statistical models for correlated and clustered data (the generalised estimating equation) were applied. In detail, the generalised estimating equation on binary data were applied to consider within-centre variability and correlated data (within patient) in the per-patient analysis (see details in the appendix p 4).

In the per-lesion analysis (where there was no patient matching between the two types of reading), tests and methods for independent samples were carried out. However, potential correlations could not be excluded since major parts of the lesions were detected in the same patients. Hence, ad hoc tests and models for correlated data were applied to allow a comprehensive evaluation. In addition, the lesion detection rate (defined as the ratio of the detected lesions of the reading mode by analysis over the detected lesions by the gold standard) was also computed in the per-lesion analysis. For the evaluation of the non-inferiority test of the difference of correlated proportions, the same approach described by Liu and colleagues and Nam was adopted (see details in the appendix pp 3–4).[23,24] The comparisons of the metrics of two reading types versus the board (as gold standard) were computed accordingly with the approach originally described by McNemar and colleagues.[25]

We used the functions *tab.paired* and *sesp.mcnemar* of the R package DTComPair that combine the two tables in S4-A (and in S4-B for P2 lesions) for obtaining a multiple contingency table on which to apply the McNemar test to for comparing the main metrics (sensitivity, specificity, positive predictive value, and negative predictive value) of standard reading versus AI-assisted reading compared with the gold standard.

Reading time comparison was performed by tests for both independent (Mann-Whitney-Wilcoxon test) and paired groups (Wilcoxon rank test) using mean, SD, median, and IQR values reported as min (using decimal numbers).

Sample size calculation was assessed at per-patient analysis based on previous evidence by Ding and colleagues in which the detection rate of AI-assisted reading was 66%, compared with 49% in conventional reading.[1] Considering a two-sample non-inferiority one side test with a margin d of 0·01, type I error α of 0·025, type II error β of 80%, the sample size resulted to be 115 per group (deep convolutional neural network and control group) and was raised to 126 considering a 10% drop-out rate. Details on the sample size and power evaluation of the study are reported in the appendix (pp 2–3).

The analyses were carried out by the software R Core Team.[26] For the computation of the McNemar test, the R-packages {stat} and {exact2x2} were used. The R-package {DTCOMPair} was adopted for the comparison of the metrics of two readings versus board (as gold standard). The generalised estimating equation models were performed by R-package {geepack} and the relative risk of the comparisons of AI-assisted versus standard, standard versus board, and AI versus board readings were obtained as an approximation of the odds ratio (obtained from the generalised estimating equation model for binary outcomes), by following the method reported by Zhang and colleagues.[27] The significance level was set at p=0·05, and the CI was set at 95%. Details of sample size computation and justification are provided in the appendix (p 2).

This study was approved by the ethics committees (EC number NP4350), has been registered on

	Patients (N=133)
Age, years	66·5 (14·4)
Sex*	
Female	73 (55%)
Male	60 (45%)
Country	
France	18 (14%)
Germany	11 (8%)
Hungary	22 (17%)
Italy	35 (26%)
Spain	13 (10%)
Sweden	17 (13%)
UK	17 (13%)
Bowel cleansing†	
Excellent–good	98 (74%)
Fair–poor	35 (26%)
Main diagnosis‡	
Negative	28 (21%)
Female	21 (75%)
Male	7 (25%)
Vascular lesion	86 (65%)
Female	41 (48%)
Male	45 (52%)
Mucosal lesion	16 (12%)
Female	9 (56%)
Male	7 (44%)
Protruding lesion	3 (2·3%)
Female	2 (67%)
Male	1 (33%)

Data are mean (SD) or n (%). *As defined in the participants' electronic health record registration. †Bowel cleansing in the standard reading. ‡Lesions defined by the board: vascular lesion (angioectasia, blood or clot, red spot, or venous angioma); mucosal lesion (erythematous mucosa, erosion, or ulcer); protruding lesion (polyp, tumour).

Table 1: **Baseline characteristics**

ClinicalTrials.gov (NCT04821349), and follows the CONSORT-AI and SPIRIT-AI extensions, as well as the EASE-SAGER guidelines (the protocol and guideline checklists are available in the appendix [pp 16–35]).

Role of the funding source

The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Results

From Feb 17, 2021 to Dec 29, 2021, 138 patients were screened; one patient was excluded due to SARS-CoV-2; thus, 137 patients were enrolled (figure 1). Four patients dropped out (two patients underwent capsule endoscopy but documentation was not provided completely, one patient was not able to swallow, and one patient underwent capsule endoscopy but the ProScan function had a technical failure which prevented the AI-assisted reading). 133 patients (73 [55%] were female, mean age 66·5 years [SD 14·4]) were analysed (table 1). Capsule endoscopy was completed in 112 (84%) of 133 patients. Bowel cleansing was adequate in 73·7% (standard readers) and 77·4% (AI-assisted readers) of patients (McNemar p=0·30).

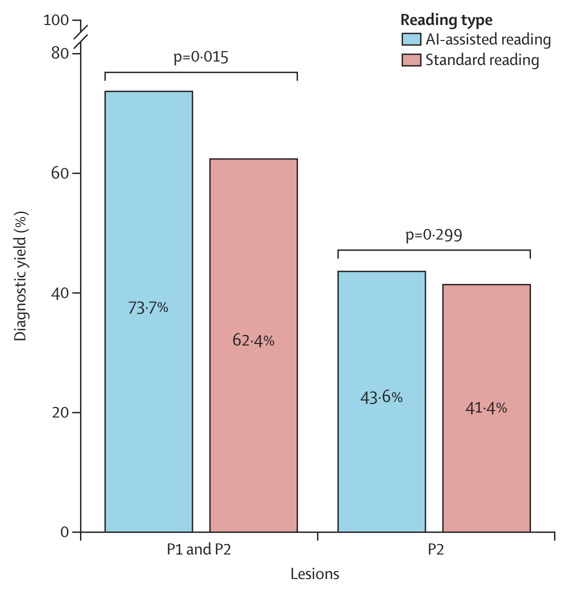


Figure 2: **Comparison of diagnostic yields of standard and AI-assisted readings at per-patient analysis**

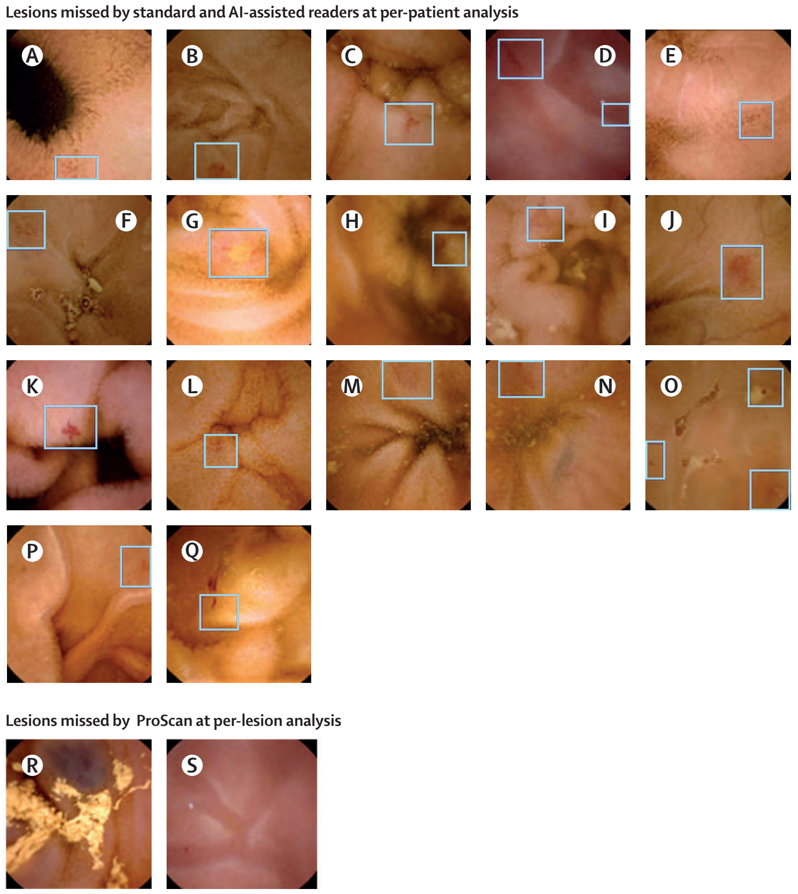
p values refer to independent sample non-inferiority tests. The results were confirmed by non-inferiority test for paired data and by the generalised estimating equation model for correlated data (appendix p 9). AI=artificial intelligence.

P1 and P2 lesions were identified in 83 of 133 patients (standard reading, diagnostic yield 62·4%) and 98 of 133 patients (AI-assisted reading, diagnostic yield 73·7%). The board reading identified P1 and P2 lesions in 105 of 133 patients (diagnostic yield 78·9%; appendix p 8). Comparing the diagnostic yield of standard and AI-assisted reading, an 11·3% increase (95% CI 3·6–19·0) was observed. Since the CI did not cross the non-inferiority margin (difference of –1%) nor the superiority margin of 1%, the diagnostic yield of AI-assisted reading was proven non-inferior and superior to standard reading (McNemar non-inferiority p<0·0001; McNemar superiority p=0·0213; figure 2). Compared with the board, reading in standard mode significantly differed (–16·5% [95% CI –23 to –10]; McNemar p<0·0001), whereas AI-assisted reading revealed no difference (–5·2% [95% CI –10·9 to 0·4]; McNemar p=0·060). Thus, the diagnostic yield of AI-assisted reading was not statistically different from that of the board. These findings were confirmed also by the results of the generalised estimating equation models and the computation of the corresponding relative risks reported in the appendix (p 9). Sensitivity, specificity, positive predictive value, and negative predictive value of standard

	P1 and P2 lesions			P2 lesions		
	Standard reading	AI assisted reading	p value*	Standard reading	AI assisted reading	p value*
Sensitivity	79·0	93·3	0·0052	84·6	89·2	0·60
Specificity	100·0	100·0	1	100·0	100·0	1
Positive predictive value	100·0	100·0	1	100·0	100·0	1
Negative predictive value	56·0	80·0	0·0303	87·2	90·7	0·65
Diagnostic accuracy	83·5	94·7	0·0056	92·5	94·7	0·52

Per-patient analysis. Contingency tables are available in the appendix (p 8). *p values were carried out by DTComPair R-package following the method proposed by McNemar.[25]

Table 2: **Diagnostic performance metrics of AI-assisted and standard reading (with board reading as the gold standard)**


Lesions missed by standard and AI-assisted readers at per-patient analysis

Lesions missed by ProScan at per-lesion analysis

Figure 3: **Lesions missed by standard and AI-assisted readers at per-patient analysis and by ProScan at per-lesion analysis**

Significant P2 lesions missed by standard (A–J) and AI-assisted (K–Q) readers. All lesions in A–Q have the blue box because ProScan detected all of them. Angiectasia: A, B, C, D, E, F, H, I, J, K, L, M, N, and P. Ulcer: G and O. Clot: Q. Venous angioma: R. Red spot: S.

and AI-assisted readings are shown in table 2 and the appendix (p 8). In comparison with the board, standard readers missed P1 and P2 lesions in 22 patients (ten P2 lesions; 12 P1 lesions), whereas AI-assisted readers missed P1 and P2 lesions in seven patients (all P2 lesions; figure 3). The miss rate for P1 and P2 lesions by standard readers (22 [21·0%] of 105) was significantly higher than that of AI-assisted readers (7 [6·6%] of 105; McNemar p=0·0094). After the adjudication, all P1 and P2 lesions missed by standard and AI-assisted readers were confirmed and reclassified as a misinterpretation of readers who failed to detect the lesions. The ProScan first automated reading detected and included all the lesions in the AI video, including those missed by the AI readers, showing a sensitivity of 100%. Comparison between AI-assisted and standard reading (used as reference) is discussed in the appendix (p 10).

P2 lesions were identified in 55 of 133 patients with standard reading (diagnostic yield 41·4%) and 58 of 133 patients with AI-assisted reading (diagnostic yield 43·6%). The board of experts identified P2 lesions in 65 of 133 patients (diagnostic yield 48·9%). Comparing the diagnostic yield of standard and AI-assisted readings, it was not possible to confirm the non-inferiority of AI-assisted reading since the 95% CI (–3·8 to 8·3) of the diagnostic yield difference included the non-inferiority margin of –1% (figure 2). When the diagnostic yields of standard and AI-assisted reading were compared with the diagnostic yield of the board, a –7·5% (95% CI –15 to 0·3; standard reading) and –5·3% (–9 to 1; AI-assisted reading) difference was observed, resulting in no difference for both the comparisons. These finding were also confirmed by the results of the generalised estimating equation models (appendix p 9). Accuracy of standard and AI-assisted readings are shown in table 2. The miss rate of standard readers (10 [15·4%] of 65) was not statistically higher than that of AI-assisted readers (7 [10·8%] of 65; McNemar p=0·623). Comparison between the AI-assisted and standard readings (used as the reference) in the detection of P2 lesions is reported in the appendix (p 10).

The board identified 364 P1 and P2 lesions (n=241 P1; n=123 P2) in 105 patients. The remaining 28 patients had a negative examination result. Standard readers identified 236 P1 and P2 lesions (n=138 P1; n=98 P2) in 83 patients (236 of 364 lesions, lesion detection rate of 64·8%). AI-assisted readers identified 320 P1 and P2 lesions (n=212 P1; n=108 P2) in 98 patients (320 of 364 lesions, lesion detection rate of 87·9%; appendix pp 11, 14). Comparing the lesion detection rate of P1 and P2 lesions of standard and AI-assisted readers, a 23·1% (95% CI 16·9–29·3) difference was observed. Since the CI did not cross the non-inferiority margin d of –1% nor the superiority margin of 1%, the lesion detection rate for P1 and P2 lesions of AI-assisted readings was non-inferior and superior to standard reading (independent sample proportion test p<0·0001; McNemar p<0·0001). These findings were confirmed also by the results of the generalised estimating equation model and the computation of the corresponding relative risk reported in the appendix (p 9). In comparison to the board reading, standard readers missed 128 lesions (n=103 P1; n=25 P2) and AI-assisted readers missed 44 lesions (n=29 P1; n=15 P2; appendix pp 11). The miss rate for P1 and P2 lesions of standard readers (n=128/364) was higher than that of AI-assisted readers (n=44/364, 35·2% *vs* 12·1%, independent sample and McNemar both p<0·0001). During the adjudication, 172 discordant lesions were evaluated (128 lesions were missed by standard readers and detected by AI [n=103 P1; n=25 P2] and 44 lesions were detected by standard readers and missed by AI [n=29 P1; n=15 P2]). All 128 lesions missed by standard readers were confirmed by the board and reclassified as a misinterpretation of the readers who missed the lesions. Similarly, 42 out of 44 lesions missed by AI-assisted readers were confirmed to be present in the AI-video but undetected by AI-assisted readers. The remaining two lesions (an ileal haemangioma and a jejunal red spot) were missed by the ProScan first automated reading (figure 3). The sensitivity of the automated ProScan reading for P1 and P2 lesions was 99·5% (362 of 364 lesions). The comparison between AI-assisted and standard reading (used as the reference) for P1 and P2 lesions is reported in the appendix (p 12).

Looking at P2 lesions exclusively, the lesion detection rate was 79·7% (98 of 123 lesions) with standard reading and 87·8% (108 of 123 lesions) with AI-assisted reading. AI-assisted reading was non-inferior (independent sample proportion test p value p=0·026, McNemar p<0·0001) to standard reading (the 95% CI of the lesion detection rate difference was –0·8 to 16·1, non-inferiority margin of –1% not included; appendix p 11). No difference was found in the miss rate of standard reading (25 [20·3%] of 123 lesions) and AI-assisted reading (15 [12·2%] of 123 lesions; independent sample proportion test p=0·120; McNemar p=0·200). The sensitivity of ProScan for P2 lesions was 122 (99·2%) of 123 lesions. Comparison between AI-assisted and standard reading (used as the reference) for P2 lesions is reported in the appendix (p 12).

Mean reading times for the small bowel and the entire video are reported in table 3. The average time to detect one lesion was 18·7 min for standard readers and 1·6 min for AI-assisted readers, with the average number of lesions detected per-patient by readers in the two modalities equal to 1·8 (standard) and 2·4 (AI-assisted). The ProScan system allowed a 20-times reduction in the mean number of images composing the videos, from 28 810 (SD 18 825) to 1199 (1972) images.

	Standard reading	AI-assisted reading	p value*
Small bowel	33·7 (22·9); 28 (20·6–40·0)	3·8 (3·3); 3 (1·4–5·0)	<0·0001
All gastrointestinal tract	43·5 (25·9); 37·75 (30·2–50·6)	6·0 (4·9); 5 (3·2–8·0)	<0·0001
Time to detect one small bowel lesion	18·7	1·6	<0·0001

Data are mean (SD) min or median (IQR) min. *p values were computed by Wilcoxon rank test.

Table 3: **Reading time of AI-assisted and standard reading**

Discussion

This study shows the non-inferiority and the superiority of AI-assisted reading versus standard reading in the detection of potentially small bowel bleeding lesions at both per-patient and per-lesion analysis, with a significantly reduced reading time, from 33·7 min to 3·8 min. To our knowledge, this is the first prospective multicentre trial evaluating the performance of AI-assisted (ProScan, Ankon, China, and AnX Robotica, USA) reading in small bowel capsule endoscopy in a real-world setting, using entire small bowel capsule endoscopy videos. Several highlights should be emphasised. First, AI-assisted reading, when considering overall P1 and P2 lesions, both at per-patient and per-lesion analysis, was confirmed to be non-inferior and superior to standard reading. P0 lesions were intentionally not considered since they do not impact clinical outcome. When limiting the analysis to P2 lesions, AI-assisted reading was confirmed to be non-inferior to standard reading at per-lesion analysis only. Although this might be considered a limitation of the system that seems to be unable to offer an adjunctive help and to assist the physicians when dealing with the most relevant findings, the results confirm that AI could act as a tool to help the reader focus on small, fleeting lesions that might easily be missed. This is further confirmed by the higher diagnostic yield which improved with AI-assisted reading at per-patient analysis. Second, AI-assisted reading significantly exceeds the accuracy of standard reading in terms of sensitivity, negative predictive value, and diagnostic accuracy, and specificity and positive predictive value were both 100%. The excellent AI-assisted reading sensitivity, as well as the higher diagnostic accuracy at per-patient analysis for P1 and P2 lesions, reassures about the minimised risk of missing lesions by the AI system and provides an estimation of the overall high accuracy and efficiency of AI-assisted reading. Third, a board of experts was used to reduce bias related to reader's misinterpretation. The board was also used for adjudication in case of discrepancies. The adjudication process provides an estimation of the software performance since the noise related to the reader's variability or misinterpretation is minimised. The results confirmed the excellent sensitivity previously reported in the validation studies, with 100% sensitivity for P1 and P2 lesions at per-patient analysis, and a sensitivity of 99·5% (for P1 and P2 lesions) and 99·2% (for P2 lesions) at per-lesion analysis.[19,28] Fourth, AI-assisted reading reduces the miss rate, especially for findings that are less visible, such as P1 lesions. At per-patient and per-lesion analysis, the miss rate for P1 and P2 lesions was lower with AI-assisted readers than with standard readers. These results are in line with those measured by Xie and colleagues who described a lesion miss rate of 4·1% for AI-assisted readers and 23·9% for standard readers.[28] Finally, the first ProScan automated reading results in a significant reduction of reading time. Mean small bowel reading time was reduced by almost nine times, confirming the results previously reported.[19,28,29]

Our study has some limitations. From a procedural standpoint, the assessment of inter-reader variability was not conducted. Nevertheless, the readers adhered to rigorous and standardised assessment procedures. Second, the sample size was not very large. However, the demographic features of our sample exhibited consistency with those documented in recent literature, ensuring comparison of our findings.[30–32] With respect to clinical limitations, deep enteroscopy was not performed routinely to confirm or exclude the presence of findings. It can, therefore, be argued that it is not possible to exclude that both reading modalities could have missed or misinterpreted findings. However, the reading was reiterated and performed also by a board of experts, neither is it ethically acceptable to perform deep enteroscopy in patients with no relevant findings. Clinical outcomes were not evaluated. Moreover, the study addressed only patients with suspected small bowel bleeding, focusing the evaluation of AI performance mainly to angioectasias, which represent the most common findings in this setting. Other pathological conditions such as Crohn's disease, hereditary polyposis syndromes, and suspected small bowel tumour, which present with more challenging small bowel capsule endoscopy findings (ie, scattered erosions, polyps, or a single submucosal mass) were not evaluated. However, the present study was primarily designed to test the accuracy and the reading time of deep convolutional neural network-assisted reading when used as a first reader to support physicians with the most common indication of small bowel capsule endoscopy which is suspected small bowel bleeding. Further studies specifically designed to define the impact of an AI-based reading on the final patient outcome, as well as on its diagnostic accuracy for any indication of small bowel capsule endoscopy, are needed. With respect to the adopted classification (Saurin classification), it could be considered not sufficiently updated according to the recent evidence. As an example, although mucosal red spots in the Saurin classification are considered P1 lesions, nowadays they are deemed clinically irrelevant. In this perspective, the clinical gain of AI-assisted reading that increased the detection of P1 lesions is not assessable and might be assessed in future trials by integrating small bowel capsule endoscopy results with the outcomes obtained at enteroscopy. Moreover, AI-based systems are not aimed to replace the reader. They are indeed developed to support the physician who is still responsible for discriminating the clinical relevance of detected lesions and making the final decision.

Contributors

CS conceptualised, designed, and supervised the study, accessed all the raw data sets, validated the study data, and prepared the draft manuscript. SP contributed to the supervision of the study, accessed all the raw data sets, validated the study data, participated in data curation and verification, did the formal data analysis, and prepared the draft manuscript. CH interpreted the results, and participated in reviewing and editing of the manuscript. CF accessed all the raw data sets, performed formal statistical analysis, interpreted the results, and prepared the draft manuscript. TN supervised and administrated the project, accessed all the raw data sets, participated in data curation and verification, and prepared the draft manuscript. BG-S, MK, and EJD accessed all the raw data sets, and validated the study data. ET, MM, ÁF, AR, XD, MER, RB, AC, AH, J-CS, AM, GWJ, AG, PB, RS, MS, KH, and AN performed small bowel capsule endoscopy readings in standard or AI-assisted reading, reviewed and verified the data, and participated in reviewing and editing. RL, GC, and DS reviewed and verified the data, and participated in reviewing and editing. All authors were responsible for the decision to submit the manuscript.

Declaration of interests

CS, SP, ET, MK, XD, and EJD received support for attending meetings or travel from AnX Robotica. CS, MK, and EJD received honoraria from AnX Robotica for the adjudication committee. J-CS received teaching fees from AnX Robotica. RS participated in an Ankon-X-advisory meeting. TN received consulting fees from AnX Robotica. AnX Robotica provided all centres with the Navicam SB system at no cost, in instances where it was not readily available. Complimentary capsules were supplied in centres where regional reimbursement did not encompass the procedure.

Data sharing

All data and source code can be made available after the manuscript publication on reasonable request, which must include an appropriate protocol, analysis plan, and data exchange with institutional approvals in place before data transfer of any information.

Acknowledgments

We sincerely thank all patients and medical and non-medical staff across the participating centres for their invaluable contribution. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. ANKON Technologies, China and AnX Robotica, USA provided the NaviCam SB system.

References

1 Pennazio M, Spada C, Eliakim R, et al. Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European Society of Gastrointestinal Endoscopy (ESGE) Clinical Guideline. *Endoscopy* 2015; **47:** 352–76.

2 Liao Z, Gao R, Xu C, Li ZS. Indications and detection, completion, and retention rates of small-bowel capsule endoscopy: a systematic review. *Gastrointest Endosc* 2010; **71:** 280–86.

3 Cortegoso Valdivia P, Skonieczna-Żydecka K, Elosua A, et al. Indications, detection, completion and retention rates of capsule endoscopy in two decades of use: a systematic review and meta-analysis. *Diagnostics (Basel)* 2022; **12:** 1105.

4 Leenhardt R, Koulaouzidis A, Histace A, et al. Key research questions for implementation of artificial intelligence in capsule endoscopy. *Therap Adv Gastroenterol* 2022; **15:** 17562848221132683.

5 Messmann H, Bisschops R, Antonelli G et al. Expected value of artificial intelligence in gastrointestinal endoscopy: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. *Endoscopy* 2022; **54:** 1211–31.

6 Yung DE, Sykes C, Koulaouzidis A. The validity of suspected blood indicator software in capsule endoscopy: a systematic review and meta-analysis. *Expert Rev Gastroenterol Hepatol* 2017; **11:** 43–51.

7 Beg S, Wronska E, Araujo I, et al. Use of rapid reading software to reduce capsule endoscopy reading times while maintaining accuracy. *Gastrointest Endosc* 2020; **91:** 1322–27.

8 Piccirelli S, Mussetto A, Bellumat A, et al. New generation express view: an artificial intelligence software effectively reduces capsule endoscopy reading times. *Diagnostics (Basel)* 2022; **12:** 1783.

9 Leenhardt R, Vasseur P, Li C, et al. A neural network algorithm for detection of GI angiectasia during small-bowel capsule endoscopy. *Gastrointest Endosc* 2019; **89:** 189–94.

10 Tsuboi A, Oka S, Aoyama K, et al. Artificial intelligence using a convolutional neural network for automatic detection of small-bowel angioectasia in capsule endoscopy images. *Dig Endosc* 2020; **32:** 382–90.

11 Klang E, Barash Y, Margalit RY, et al. Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy. *Gastrointest Endosc* 2020; **91:** 606–13.e2.

12 Aoki T, Yamada A, Kato Y, et al. Automatic detection of various abnormalities in capsule endoscopy videos by a deep learning-based system: a multicenter study. *Gastrointest Endosc* 2021; **93:** 165–73.e1.

13 Ribeiro T, Saraiva MM, Ferreira JPS, et al. Artificial intelligence and capsule endoscopy: automatic detection of vascular lesions using a convolutional neural network. *Ann Gastroenterol* 2021; **34:** 820–28.

14 Wang S, Xing Y, Zhang L, Gao H, Zhang H. Deep convolutional neural network for ulcer recognition in wireless capsule endoscopy: experimental feasibility and optimization. *Comput Math Methods Med* 2019; **2019:** 7546215.

15 Mascarenhas Saraiva MJ, Afonso J, Ribeiro T, et al. Deep learning and capsule endoscopy: automatic identification and differentiation of small bowel lesions with distinct haemorrhagic potential using a convolutional neural network. *BMJ Open Gastroenterol* 2021; **8:** e000753.

16 Mohan BP, Khan SR, Kassab LL, et al. High pooled performance of convolutional neural networks in computer-aided diagnosis of GI ulcers and/or hemorrhage on wireless capsule endoscopy images: a systematic review and meta-analysis. *Gastrointest Endosc* 2021; **93:** 356–64.e4.

17 Hosoe N, Horie T, Tojo A, et al. Development of a deep-learning algorithm for small bowel-lesion detection and a study of the improvement in the false-positive rate. *J Clin Med* 2022; **11:** 3682.

18 Hwang Y, Lee HH, Park C, et al. Improved classification and localization approach to small bowel capsule endoscopy using convolutional neural network. *Dig Endosc* 2021; **33:** 598–607.

19 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:** 1044–54.e5.

20 Rondonotti E, Spada C, Adler S, et al. Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European Society of Gastrointestinal Endoscopy (ESGE) Technical Review. *Endoscopy* 2018; **50:** 423–46.

21 Saurin JC, Delvaux M, Gaudin JL, et al. Diagnostic value of endoscopic capsule in patients with obscure digestive bleeding: blinded comparison with video push-enteroscopy. *Endoscopy* 2003; **35:** 576–84.

22 Brotz C, Nandi N, Conn M et al. A validation study of 3 grading systems to evaluate small-bowel cleansing for wireless capsule endoscopy: a quantitative index, a qualitative evaluation, and an overall adequacy assessment. *Gastrointest Endosc* 2009; **69:** 262–70.

23 Liu J, Hsueh H, Hsieh E, Chen JJ. Tests for equivalence or non-inferiority for paired binary data. *Stat Med* 2002; **21:** 231–45.

24 Nam JM. Establishing equivalence of two treatments and sample size requirements in matched-pairs design. *Biometrics* 1997; **53:** 1422–30.

25 McNEMAR Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; **12:** 153–57.

26 R Core Team. R: a language and environment for statistical computing. 2023. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

27 Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 1998; **280:** 1690–91.

28 Xie X, Xiao YF, Zhao XY, et al. Development and validation of an artificial intelligence model for small bowel capsule endoscopy video review. *JAMA Netw Open* 2022; **5:** e2221992.

29 Ding Z, Shi H, Zhang H, et al. Artificial intelligence-based diagnosis of abnormalities in small-bowel capsule endoscopy. *Endoscopy* 2023; **55:** 44–51.

30 Brito HP, Ribeiro IB, de Moura DTH, et al. Video capsule endoscopy *vs* double-balloon enteroscopy in the diagnosis of small bowel bleeding: a systematic review and meta-analysis. *World J Gastrointest Endosc* 2018; **10:** 400–21.

31 Girelli CM, Soncini M, Rondonotti E. Implications of small-bowel transit time in the detection rate of capsule endoscopy: a multivariable multicenter study of patients with obscure gastrointestinal bleeding. *World J Gastroenterol* 2017; **23:** 697–702.

32 Elli L, Scaramella L, Tontini GE, et al. Clinical impact of videocapsule and double balloon enteroscopy on small bowel bleeding: results from a large monocentric cohort in the last 19 years. *Dig Liver Dis* 2022; **54:** 251–57.