87481_Auto_Edited.docx

WORD COUNT	2766	TIME SUBMITTED	21-NOV-2023 08:26AM
		PAPER ID	104410851

Name of Journal: *Artificial Intelligence in Gastrointestinal Endoscopy*

Manuscript NO: 87481

Manuscript Type: ORIGINAL ARTICLE

Retrospective Cohort Study

Artificial intelligence fails to improve colonoscopy quality: A single centre retrospective cohort study

Naeman Goetz, Katherine Hanigan, Richard Kai-Yuan Cheng

Abstract

BACKGROUND

Limited data currently exists on the clinical utility of Artificial Intelligence Assisted Colonoscopy (AIAC) outside of clinical trials.

AIM

To evaluate the impact of AIAC on key markers of colonoscopy quality compared to conventional colonoscopy (CC).

METHODS

This single-centre retrospective observational cohort study included all patients undergoing colonoscopy at a secondary centre in Brisbane, Australia. CC outcomes between October 2021 and October 2022 were compared with AIAC outcomes after the introduction of the Olympus Endo-AID module from October 2022 to January 2023. Endoscopists who conducted over 50 procedures before and after AIAC introduction were included. Procedures for surveillance of inflammatory bowel disease were excluded. Patient demographics, proceduralist specialisation, indication for colonoscopy, and colonoscopy quality metrics were collected. Adenoma detection rate (ADR) and sessile serrated lesion detection rate (SSLDR) were calculated for both AIAC and CC.

RESULTS

The study included 746 AIAC procedures and 2162 CC procedures performed by seven endoscopists. Baseline patient demographics were similar, with median age of 60 years with a slight female predominance (52.1%). Procedure indications, bowel preparation quality, and caecal intubation rates were comparable between groups. AIAC had a slightly longer withdrawal time compared to CC, but the difference was not statistically significant. The introduction of AIAC did not significantly change ADR (52.1% for AIAC *vs* 52.6% for CC, $P = 0.91$) or SSLDR (17.4% for AIAC *vs* 18.1% for CC, $P = 0.44$).

CONCLUSION

The implementation of AIAC failed to improve key markers of colonoscopy quality, including ADR, SSLDR and withdrawal time. Further research is required to assess the utility and cost-efficiency of AIAC for high performing endoscopists.

INTRODUCTION

Screening colonoscopy has been instrumental in reducing the incidence and mortality from colorectal cancer (CRC). However, up to 9% of CRCs develop in patients up-to-date on surveillance colonoscopies, termed interval cancers, thought to overwhelmingly result from suboptimal examination[1]. In defining colonoscopy quality, the most widely used quality metric is adenoma detection rate (ADR), which is the proportion of screening colonoscopies where at least one adenoma is found[2]. As ADR is inversely correlated to the interval cancer rate[3], technology that can aid adenoma detection has been the focus of intense research.

Artificial Intelligence Assisted Colonoscopy (AIAC) has emerged as a potential tool for improving colonoscopy quality and mitigating factors such as proceduralist fatigue or inattention in a procedure that is substantially operator dependent. Early robust randomised controlled trial (RCT) data on computer-aided polyp detection (CADe),

which involves neural networks processing colonoscopy images in real time and superimposing a visual alert over suspected polyps on the endoscopy display, has garnered strong enthusiasm for this field[4]. Indeed, meta-analysis of published RCTs suggest that CADe can improve ADR by as much as 10%[5,6]. However, the majority of included trials were single-center studies conducted largely in Chinese institutions with relatively low baseline ADRs and using proprietary technology not available commercially[7]. As such, published data on the utility and cost-effectiveness in real-world clinical settings is limited. The objective of our study was to assess the effect of AIAC on key benchmarks of colonoscopy quality including ADR, sessile serrated lesion detection rate (SSLDR), and withdrawal time in comparison to conventional colonoscopy (CC).

MATERIALS AND METHODS

This was a single-centre retrospective observational cohort study conducted at Redcliffe Hospital, a public secondary hospital in Brisbane, Australia, which provides an open-access endoscopy service. All consecutive colonoscopies from October 2021 until January 2023 were included in the study. Patients were identified through a prospectively maintained departmental database of all patients undergoing colonoscopy. The introduction of the Olympus End-AID module in October 2022 allowed us to compared outcomes for CC in the preceding year with those of AIAC in the subsequent three months. For inclusion, proceduralists must have performed at least 50 colonoscopies both before and after the introduction of the Endo-AID module. We also only included patients with an intact colon. Colonoscopies performed for the surveillance of inflammatory bowel disease were excluded. All endoscopists included in the study had at least five years of independent endoscopy experience.

The primary endpoint for the study was change in three surrogate markers of colonoscopy quality with artificial intelligence (AI): ADR, SSLDR and withdrawal time. Additional variables collected included patient demographics, proceduralists

specialisation, indication for colonoscopy, polyp size and colonoscopy quality metrics including bowel preparation and caecal intubation rate.

Proceduralists were able to switch the AI assistance mode on and off and use adjunctive techniques to enhance polyp detection, such as distal cap, narrow band imaging or chemical chromoendoscopy at their discretion. All patients underwent split bowel preparation, and the quality of preparation was evaluated and graded using the Boston Bowel Preparation Scale by the performing proceduralist. All procedures were performed under conscious sedation using a combination of fentanyl and midazolam. The final decision regarding polyp resection was at the discretion of the proceduralist. Procedures were conducted in one of two dedicated endoscopy rooms equipped with identical high-definition colonoscopes (Olympus EVIS EXTRA) and histopathology was performed at a single laboratory, Queensland Pathology.

Statistical analysis

Statistical analysis was performed using Stata Corp STATA software (Boston, United States). Using ADR as the primary outcome and anticipating an effect size of 0.10, we calculated a minimum sample size of 236 patients to achieve a 95% confidence interval. Univariate comparisons of baseline parameters were conducted using the unpaired *t*-test after confirming normal distribution. Non-parametric data was assessed using the Mann-Whitney U test, while categorical data was analysed using the Chi-squared or Fisher's exact test. We set statistical significance at a *P* value of < 0.05.

RESULTS

We compared 746 AIACs with 2126 CCs, which were conducted by seven endoscopists, comprising four gastroenterologists and three surgeons. Patient demographics were similar between patients undergoing AIAC and CC at baseline, with a median age of 60 years [interquartile range (IQR) 49-70] and a slight female predominance of 52.1% (Table 1). Procedure indications in order of frequency were symptoms (35.1%), surveillance following previous polyps (31.2%) and investigation of a positive faecal occult blood test

(14.8%). The indication for the procedure, quality of bowel prep and caecal intubation rates were well matched between the study populations ($P > 0.05$).

AIAC introduction ultimately had no significant impact on either ADR (52.1% for AIAC *vs* 52.6% for CC, $P = 0.91$) or SSLDR (17.4% for AIAC *vs* 18.1% for CC, $P = 0.44$) on an institutional level. However, a per-proceduralist analysis (Figure 1) demonstrated a significant change for two endoscopists, with ADR increasing by 16.8% for one (CC 61.4%, AIAC 78.2%, $P = 0.004$) and decreasing by 21% for another (CC 58.6%, AIAC 37.6%, $P = 0.006$). By-proceduralist analysis of SSLDR did not yield significant results. The AIAC group exhibited a longer mean withdrawal time (13 min 18 sec) compared with the CC group (12 min 29 sec), though this different was not statistically significant ($P = 0.48$) (Figure 2). Analysis by adenoma or sessile serrated lesion (SSL) size was not significant between groups, with the majority of adenomas detected being < 5 mm in size in both groups (Figure 3).

DISCUSSION

We demonstrate that AI in colonoscopy yielded no benefit in our unit and failed to improve either ADR or SSLDR. One possible explanation for our experience being discordant to trial data is that the baseline ADR of 52% in our unit is substantially higher than many of the published RCTs to date, and there may be a ceiling effect to polyp detection among high performing endoscopists. Furthermore, results from RCTs may be overly optimistic as proceduralists were not blinded to the intervention, which may have impacted their performance and prompted a more thorough mucosal exposure or conscientious lesion assessment[8]. A more concerning explanation would be that CADe instilled a false sense of security and unwittingly resulted in a degradation of mucosal exposure quality, though a consistent withdrawal time would argue against this. It is also possible that proceduralists did not utilize CADe to the full extent and ignored lesions highlighted by CADe because they either deemed these to be clinically unimportant or incorrectly believed them to be false-positive signals. As such, exploration of endoscopist

attitudes and behavior in the face of a nascent technology and formal training in CADe may be critical for successfully integrating AIAC across a range of practice settings.

Interestingly, per-proceduralist analysis of ADR yielded significant results for two individuals, including one interventional gastroenterologist whose ADR deteriorated with AIAC. Given the comparatively short period of observation of AIAC compared with CC, this may reflect a type II error due to the smaller number of AIAC procedures or alternatively stem from an altered referral pattern during this limited time period. Again, there was no change in withdrawal time to suggest a degradation in examination quality due to overreliance on AI. In terms of SSL detection, these are known to be more challenging to detect given they are often located in the proximal colon and have a non-polypoid configuration with inconspicuous borders[9]. There has been no substantial improvement in SSLDR in the majority of published AIAC studies except for two tandem colonoscopy RCTs which demonstrated reduced SSL miss rates with AIAC, although detection rates compared unfavorably with our 18% baseline SSLDR[10,11].

Our study is not an outlier, but rather follows a series of recent disappointing results from AIAC implementation in high-performing Western endoscopy units that challenge the generalizability of the benefits of CADe demonstrated in early RCTs across broader clinical settings. Most notably, Wei *et al*[12] performed a multi-center RCT across four community-based endoscopy centers in the United States and found no change in the number of adenomas per colonoscopy (0.73 for AIAC *vs* 0.67 for CC, $P = 0.496$) or the ADR (35.9% for AIAC *vs* 37.2% for CC, $P = 0.774$). This study is particularly salient as it offered a more pragmatic trial design, allowing proceduralists to choose how they employed the AI assistance mode (*i.e.* 'on' during insertion or only once the cecum was reached). Similarly, a United Kingdom RCT found that AIAC resulted in a higher polyp detection rate (85.7% for AIAC *vs* 79.7% for CC; $P = 0.05$) but no change in ADR (71.4% for AIAC *vs* 65.0% for CC, $P = 0.09$)[13]. Furthermore, a large volume endoscopy center in Israel retrospectively demonstrated a deterioration in ADR with AIAC implementation (30.3% for AIAC *vs* 35.2% for CC, $P < 0.001$)[14]. While there are challenges to comparing results of different CADe systems across various clinical settings, these studies highlight

that real-world CADe implementation without attention to the AI-human interaction may fail to achieve intended outcomes.

Withdrawal time is a key marker of colonoscopy quality that is strongly correlated with ADR[15]. In our study, withdrawal time did not change, though our median of 12.82 min for CC significantly exceeds the grouped averages for controls in early RCTs which ranged from 4.76 min to 6.99 min[6]. In these RCTs, improvements in ADR with CADe have paralleled increases in withdrawal time. Similarly, a New Zealand center demonstrated increased ADR with AIAC deployment (47.9% for AIAC *vs* 38.5% for CC; $P = 0.03$), though the AIAC group also had a significantly longer withdrawal time (15 min for AIAC *vs* 13 min for CC; $P < 0.001$)[16]. Arguably, ADR improvements could therefore merely be the result of a more thorough examination, reflected in the longer withdrawal time, rather than AI. Notably, in the study by Wei *at al*[12], ADR did not improve despite a prolonged withdrawal time in the AIAC group, possibly reflecting increased time spent assessing activations from the AI module, including possible false positives.

Even though polyp size was comparable between groups in our study, it is worth noting that improvements in ADR with AI in previous RCTs have primarily been driven by an increased in the detection of diminutive adenomas of 5 mm in size or lower[6]. In our study, adenomas < 5 mm constituted 60.7% of resected adenomas in the control group, compared with a mean of 19% in RCT controls[6]. Coupled with high baseline ADRs and SSLDRs in our unit, this likely reflects astute mucosal exposure and examination in our institution. Furthermore, the merit of increased detection and removal of diminutive polyps is a point of controversy, particularly with respect to the degree this mitigates cancer risk[17].

A significant strength of our study is its real-world setting, which confers less risk of operator bias than a trial framework. Important limitations include the retrospective design, relatively short period of observation for AIAC and lack of patient randomization, though enrollment of consecutive patients resulted in well-matched baseline characteristics. Furthermore, the CADe mode could be switched on and off by proceduralists at their discretion, generating a further variable of "on time" for CADe,

which was not documented. In addition, though intuitive, no formal training was provided for CADe prior to implementation.

CONCLUSION

Ultimately, while AIAC has shown promise in early RCTs, further validation is required to assess its effectiveness and cost-efficiency in institutions with high performance metrics at baseline, where gains from AI are likely to be far more incremental. Specifically, longitudinal studies that assess the impact of AIAC on interval cancer rates are required. Beyond CADe, additional applications of AI in colonoscopy may increase its utility. For example, the development of computer-assisted diagnosis, which promises to confidently distinguish diminutive hyperplastic polyps from neoplastic lesions through optical pathology, could lead to significant cost-savings by allowing proceduralists to adopt a 'resect and discard' policy rather than sending these specimens for histopathology[18]. Similarly, novel AI systems can recognize key endoscopic landmarks, specific tools, and quality of bowel preparation and integrate this information into an automatically generated colonoscopy report, reducing peri-procedural documentation burden[19]. As such, it may be that a comprehensive suite of AI tools is necessary to fully realize the benefits in this field.

ARTICLE HIGHLIGHTS

Research perspectives

While Artificial Intelligence Assisted Colonoscopy (AIAC) has shown promise in early randomised controlled trials (RCTs), further validation is required to assess its utility and cost-effectiveness in centres with high baseline performance metrics, where gains from artificial intelligence (AI) are likely to be far more incremental. Specifically, longitudinal studies that assess the impact of AIAC on interval cancer rates are required.

Research conclusions

In our institution, introduction of AIAC failed to improve key benchmarks of colonoscopy quality, including adenoma detection rate (ADR), sessile serrated lesion detection rate (SSLDR) and withdrawal time. An important limitation of our investigation is the relatively brief observation period following AIAC implementation, that the 'on time' of the AI assistance mode was not recorded as well as the retrospective design.

Research results

The study included 746 AIAC procedures and 2162 conventional colonoscopy (CC) procedures performed by seven endoscopists. Baseline patient demographics were similar, with a median age of 60 years and a slight female predominance (52.1%). Procedure indications, bowel preparation quality, and caecal intubation rates were comparable between groups. AIAC had a slightly longer withdrawal time compared to CC, but the difference was not statistically significant. The introduction of AIAC did not significantly change ADR (52.1% for AIAC *vs* 52.6% for CC, $P = 0.91$) or SSLDR (17.4% for AIAC *vs* 18.1% for CC, $P = 0.44$).

Research methods

This retrospective observational cohort study was conducted at a single center in Brisbane, Australia, encompassing all patients who underwent colonoscopy during the study period. Colonoscopy quality markers for CCs conducted from October 2021 to October 2022 were compared with AIAC markers following the implementation of the Olympus Endo-AID module from October 2022 to January 2023. Proceduralists who conducted over 50 procedures before and after AIAC introduction were included. Procedures for surveillance of inflammatory bowel disease were excluded. Patient demographics, proceduralist specialisation, indication for colonoscopy, and colonoscopy quality metrics were collected. We determined the ADR and SSLDR for both CC and AIAC.

Research objectives

The objective of our investigation was to assess the effect of AIAC on key benchmarks of colonoscopy quality including the detection rate of adenomas (ADR) and SLLDR as well as withdrawal time in comparison to CC.

Research motivation

In recent years, rapid technological advancements and a focus on quality improvement have garnered significant enthusiasm for AIAC as a means of improving key markers of colonoscopy quality. While early data appears promising, this technology requires validation in day-to-day clinical practice.

Research background

AIAC has emerged as a potential tool for improving colonoscopy quality and mitigating factors such as proceduralist fatigue or inattention in a procedure that is substantially operator dependent. However, published data on the utility and cost-effectiveness in real-world clinical settings is limited.

87481_Auto_Edited.docx

ORIGINALITY REPORT

4%
SIMILARITY INDEX

PRIMARY SOURCES

1	"Routine GI Endoscopy", Journal of Gastroenterology and Hepatology, 2023 Crossref	99 words — 3%
2	gastrores.org Internet	15 words — 1%
3	www.medrxiv.org Internet	12 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS