Assessment of colonoscopy skill using machine learning to measure quality: Proof-of-concept and initial validation

OPEN ACCESS

Authors

Matthew Wittbrodt[1], Matthew Klug[1], Mozziyar Etemadi[1,2], Anthony Yang[3], John E. Pandolfino[4], Rajesh N. Keswani[4]

Institutions

1 Information Services, Northwestern Medicine, Chicago, United States
2 Anesthesiology, Northwestern University Feinberg School of Medicine, Chicago, United States
3 Surgery, Indiana University School of Medicine, Indianapolis, United States
4 Medicine, Northwestern University Feinberg School of Medicine, Chicago, United States

Keywords

Quality and logistical aspects, Training, Quality management

received 22.4.2024
accepted after revision 14.5.2024
accepted manuscript online 27.5.2024

Bibliography

Endosc Int Open 2024; 12: E849–E853
DOI 10.1055/a-2333-8138
ISSN 2364-3722

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Corresponding author

Dr. Rajesh N. Keswani, Northwestern University Feinberg School of Medicine, Medicine, Chicago, United States
raj-keswani@northwestern.edu

ABSTRACT

Background and study aims Low-quality colonoscopy increases cancer risk but measuring quality remains challenging. We developed an automated, interactive assessment of colonoscopy quality (AI-CQ) using machine learning (ML).

Methods Based on quality guidelines, metrics selected for AI development included insertion time (IT), withdrawal time (WT), polyp detection rate (PDR), and polyps per colonoscopy (PPC). Two novel metrics were also developed: HQ-WT (time during withdrawal with clear image) and WT-PT (withdrawal time subtracting polypectomy time). The model was pre-trained using a self-supervised vision transformer on unlabeled colonoscopy images and then finetuned for multi-label classification on another mutually exclusive colonoscopy image dataset. A timeline of video predictions and metric calculations were presented to clinicians in addition to the raw video using a web-based application. The model was externally validated using 50 colonoscopies at a second hospital.

Results The AI-CQ accuracy to identify cecal intubation was 88%. IT ($P = 0.99$) and WT ($P = 0.99$) were highly correlated between manual and AI-CQ measurements with a median difference of 1.5 seconds and 4.5 seconds, respectively. AI-CQ PDR did not significantly differ from manual PDR (47.6% versus 45.5%, $P = 0.66$). Retroflexion was correctly identified in 95.2% and number of right colon evaluations in 100% of colonoscopies. HQ-WT was 45.9% of, and significantly correlated with ($P = 0.85$) WT time.

Conclusions An interactive AI assessment of colonoscopy skill can automatically assess quality. We propose that this tool can be utilized to rapidly identify and train providers in need of remediation.

Introduction

Although screening and surveillance colonoscopy is associated with a reduction in the risk of colorectal cancer (CRC), post-colonoscopy CRC still occurs in practice. The risk of developing cancer after colonoscopy varies based on the quality of the colonoscopist performing the examination. Although measuring colonoscopy quality metrics such as adenoma detection rate (ADR) may identify and permit intervention to reduce these variations in quality, multiple barriers to measurement exist and prevent their widespread utility. These barriers include inadequate procedure volume to confidently assess quality [1, 2], lack of resources to calculate metrics, and potential for gamification.

In previous work, we found that measuring colonoscopy skill using manual review of a small number of colonoscopy videos can serve as an estimate of colonoscopy quality metrics such as ADR, which take a significantly larger number of procedures to calculate [3]. Furthermore, assessment of skill – such as how the colonoscopist cleans the colon, looks behind folds, and distends the colon – can permit directed feedback to the colonoscopist to facilitate improvement [4]. However, manual review of colonoscopy video is laborious and suffers from interobserver variation, and thus, is not amenable to widespread implementation.

We hypothesized that machine learning (ML), which allows computer algorithms to perform tasks generally performed by humans, could assess the quality of colonoscopy skills and associated metrics in an automated fashion. Thus, the primary aim of this study was to develop and validate an automated assessment of colonoscopy inspection utilizing ML.

Methods

Setting

This study took place at two affiliated medical centers – an academic medical center and an affiliated rural hospital – both in the United States. A waiver of informed consent was obtained via the institutional review board. Videos from the academic medical center were used to develop the ML models. Videos stored from February 2022 to March 2022 were utilized during the validation phase.

Electronic health record and video storage data sources

All endoscopic reports were written in a single endoscopic reporting system (Provation, Minneapolis, Minnesota, United States) and all electronic health record data were stored in a separate system (Epic, Madison, Wisconsin, United States). All videos were stored via a commercial gastrointestinal endoscopy cloud storage company (Virgo Surgical Video Solutions, San Francisco, California, United States). Procedure videos are automatically uploaded to the cloud server. In previous work [5], our group described a process to link colonoscopy videos with provider data as well as patient demographics and outcomes.

Colonoscopy procedures

Colonoscopy procedures at the academic medical center were performed at one of two locations (16 total procedure rooms) over the study period. Colonoscopy procedures during the validation phase were performed at the rural hospital (2 total procedure rooms). During the validation phase, only colonoscopists who performed > 100 screening colonoscopies over the study period (September 1, 2018 to April 1, 2021) were included.

Definitions

A screening colonoscopy was defined as any colonoscopy performed on a patient without a personal history of colon polyps and without any gastrointestinal symptoms reported in the procedure indication. A surveillance colonoscopy was defined as a colonoscopy performed on a patient with a personal history of colon polyps without gastrointestinal symptoms reported in the procedure indication. Diagnostic procedures were procedures performed for evaluation of gastrointestinal symptoms.

Withdrawal time (WT) was defined as the duration of time spent examining the colon for colorectal polyps in procedures without polypectomy or biopsies ("normal" colonoscopies). Both the time the cecum was initially reached as well as the time the colonoscope was removed were marked by the nurse or technician. Polyp detection rate was calculated as the proportion of colonoscopies performed with removal of a polyp. Retroflexion was defined as any successful view of the endoscope and lumen in the retroflexed position; this could occur either in the right colon or rectum. The number of complete right colon evaluations was defined as the number of times the colon was inspected in its entirety from the cecum to the hepatic flexure.

Outcome measures

The primary outcome measure was the accuracy of the AI-CQ to calculate WT. Secondary outcome measures included accuracy of insertion time (IT), polyp detection rate (PDR), polyps per colonoscopy (PPC), retroflexion, and number of right colon evaluations.

We also calculated two exploratory outcome measures. WT is traditionally calculated using only normal screening colonoscopy procedures because of the infeasibility of excluding polypectomy time. To address this, we calculated WT in screening and surveillance procedures with polypectomy, automatically excluding polypectomy time (WT-PT). We calculated PT as the time from initial detection of the polyp until after the polyp was removed (i. e., no further snare resections or forceps). We also calculated high-quality WT (HQ-WT). This was defined as the amount of time in which a clear image of the colon was obtained. A clear image was based on manual labeling – only frames where the colon mucosa could be seen with clarity to identify polyps (i. e., excluding "red out", obscuring stool, or blurry image).

Model development and validation

The AI model was pre-trained using a self-supervised vision transformer on unlabeled colonoscopy images ($n = 1 \times 10^7$) mutually exclusive from all other datasets. The vision transformer model was finetuned for multi-label classification on another mutually exclusive colonoscopy image dataset (n = 9854), derived from screening, surveillance, and diagnostic colonoscopies using anatomical, procedure, and pathological labels (label n = 14). All labeling was performed by a single experienced colonoscopist (RNK).

During inference, colonoscopy video frame predictions were generated at a resolution of one frame per second and employed a binary threshold of ≥ 0.5 to denote presence; these predictions were subsequently used to calculate all metrics. A timeline of video predictions and metric calculations were presented to clinicians in addition to the raw video using a web-based application.

After model development ("AI-CQ"), the AI-CQ was externally validated using 50 screening and surveillance colonoscopies at a second affiliated hospital. All manual measurements were performed by a single experienced colonoscopist (RNK) blinded to the measurements of the AI-CQ for each video.

Statistical analysis

All data were checked for normality before analysis using the Shapiro-Wilk normality test in the stats package (v4.3.0) in R (v4.3.0). Kruskal-Wallis rank sum tests from the R stats package were employed to compare manual and AI-CQ IT and WT measurements. Spearman's rank correlations were employed to assess the association between manual and AI-CQ measurements of IT and WT. Differences in polyp detection rate were examined using Fisher's Exact Test from the R stats package. Descriptive statistics were reported using medians and interquartile range for continuous variables and percentages for categorical variables.

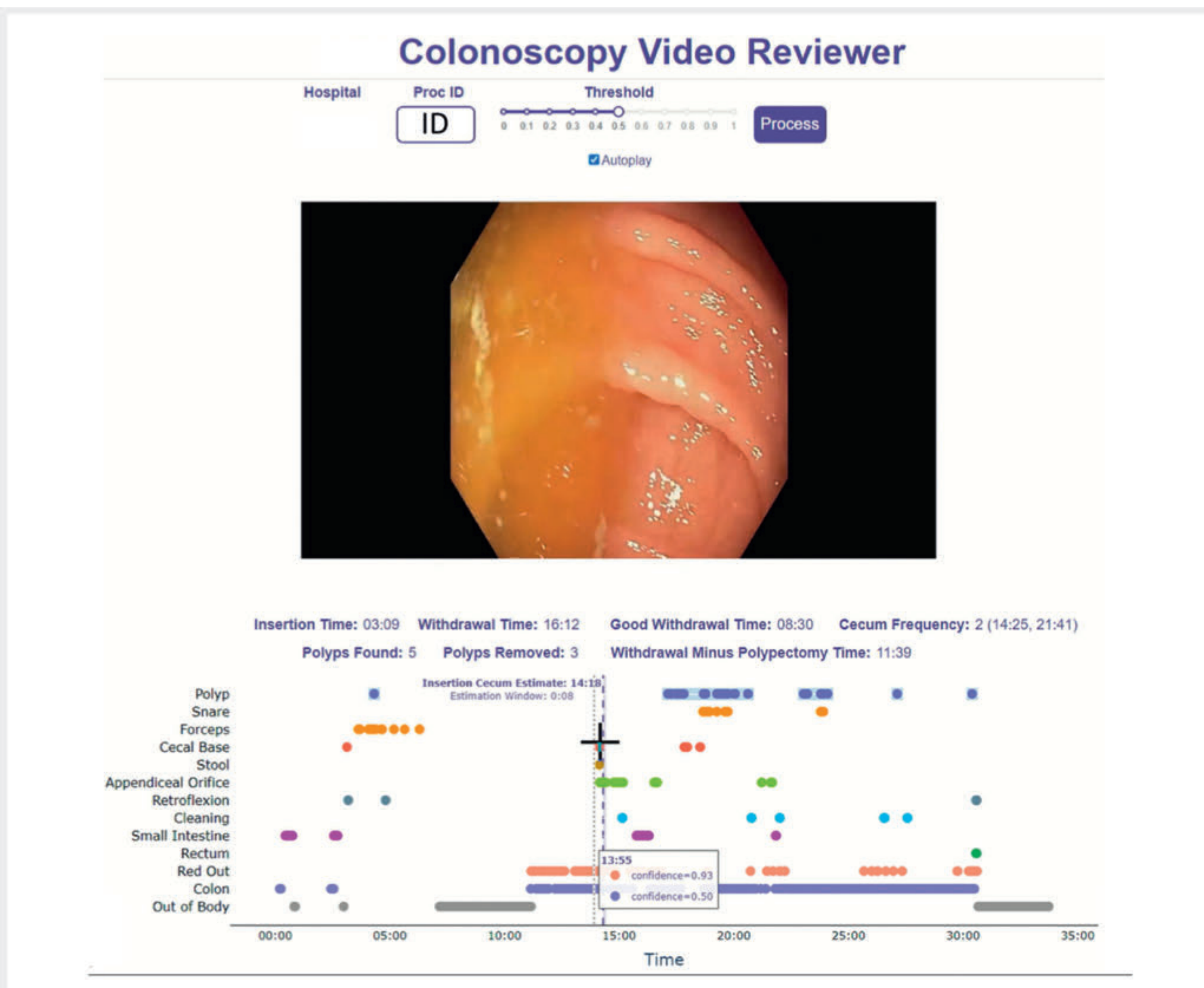
Results

The interactive AI-CQ tool is shown in ▶ **Fig. 1** and the ▶ **Video 1**. The visual tool allows the reviewer to identify relevant colonoscopy landmarks including locations outside the gastrointestinal tract, appendiceal orifice, cecal base, and small intestine; findings including polyps, stool, and unclear scope image ("red out"); devices including forceps and snares; and technical maneuvers including retroflexion, polypectomy, and cleaning.

After AI-CQ model development using videos at a single hospital, the model was externally validated using 50 screening and surveillance colonoscopy videos from six colonoscopists at a second hospital.

Cecum identification

The cecum was reached in 48 of 50 of validation cases; in the two cases in which the cecum was not reached, the AI-CQ correctly did not identify cecal intubation. Of the 48 cases in which



▶ **Fig. 1** The interface for the AI-CQ allows the user to identify relevant landmarks and maneuvers, confidence that this prediction is correct (with an adjustment bar for confidence threshold), and the ability to watch the full-length colonoscopy video. Furthermore, quality metrics predictions for the entire video are provided.

▶ **Table 1** Performance of AI-CQ tool for measuring colonoscopy quality.

	Manual	AI-CQ	Correlation
Insertion time (s)	320.5 (239)	321.5 (239.5)	$\rho = 0.99^{*}$
Median normal colonoscopy withdrawal time (s)	522 (272.5)	517.5 (270.75)	$\rho = 0.99^{*}$
Median withdrawal time – polypectomy time (s)		502 (187)	
High-quality withdrawal time (s)		237 (117)	
Polyp detection rate (%)	45.2	47.6	
Polyps per colonoscopy (mean ± SD)	0.81 ± 0.94	0.67 ± 1.1	$\rho = 0.82^{*}$

$^{*}P < 0.001$.

VIDEO

▶ **Video 1** Example of the AI-CQ utilized to identify landmarks, colonoscopy maneuvers, and calculate quality metrics. The tool allows interactive evaluation and assessment of the full-length video.

the cecum was reached, the AI-CQ correctly identified the time of cecal intubation in 88%. Of the six cases in which the cecum was reached but the AI-CQ did not identify the cecum, four were due to inadequate bowel preparation obscuring landmarks and in the remaining two, clear cecal landmarks were present but not identified. Overall, the accuracy of the AI-CQ for identifying cecal intubation was 88% (▶ **Table 1**).

Insertion and inspection time

Using cecal intubation time, IT and WT were calculated. IT (ρ = 0.99) and WT (ρ = 0.99) were highly correlated between manual and AI-CQ measurements. The median difference of calculated IT was 1.5 seconds and of WT was 4.5 seconds (▶ **Table 1**). Median HQ-WT was 45.9% (IQR: 14) of, and significantly correlated with (ρ = 0.85; $P < 0.001$), normal WT time. In colonoscopies in which a polyp was removed, median WT-PT (484 s) was similar to mean normal colonoscopy WT (502 s).

AI-CQ correctly identified rectal retroflexion in 95.2% of colonoscopies. The number of complete right colon evaluations was accurately measured in all colonoscopies. Because there is no manual method to measure the duration of cleaning, this was not validated.

Polyp detection

In aggregate, the PDR in the validation cohort was 45.2%. The AI-CQ PDR was not significantly different (47.6%, P = 0.66). The PPC in the validation cohort was 0.67; the AI-CQ measured a greater PPC (0.81; P = 0.34). In general, this occurred due to the AI-CQ counting a single polyp twice.

Discussion

Although measuring colonoscopy quality is central to CRC prevention, it remains challenging in practice. Thus, we sought to develop a proof-of-concept artificial intelligence assessment of colonoscopy quality, the AI-CQ, that automatically measures quality metrics that are traditional (e. g., WT) and more recent (e. g., number of times the right colon is fully evaluated) and identifies techniques central to high-quality colonoscopy (e. g., cleaning). Furthermore, presenting this information in an interactive application facilitates AI-augmented manual review of colonoscopy procedures. We also showed achieved initial validation that this tool performs well in measuring traditional quality metrics.

A major focus of colonoscopy AI work has been around polyp detection with multiple commercial products already approved or in development [6, 7, 8]. There has been significantly less work around developing algorithms that can measure colonoscopy quality. In an initial proof of concept, Thakkar et al described an approach that could be used to measure core colonoscopy techniques including cleaning, fold examination, and luminal distention [9]. A real-time algorithm acting as a "speedometer" to measure withdrawal speed has been described but did not improve quality [10]. In more recent work, an AI tool to measure colonoscopy WT and PT (similar to what we have described above) was described with potential added functionality of minimizing manual documentation that must occur after procedures [11].

In contrast to prior systems, the AI-CQ is meant to be an interactive tool. The tool loads a recorded colonoscopy video and analyzes it on demand for review. We propose that this interactive application can be utilized in multiple settings that have been shown to be effective in prior research but are not feasible for routine use. Potential applications would be providing feedback on withdrawal technique, similar to work we and others have previously published [3]. For example, the "expert" and learner could watch the video together with AI identifying relevant areas to focus on. In other prior work, the importance of providing feedback on polypectomy technique to both practicing colonoscopists [4] and trainees [12] has been demonstrated. However, identifying which colonoscopies have polyps removed, where in a video the polypectomy occurs, and using which tool is time-consuming. Thus, AI-augmented video review of colonoscopy quality is an opportunity to feasibly provide substantive feedback to colonoscopy trainees and those requiring remediation.

There are important limitations to this study. While all algorithms were externally validated using videos from a second site, all videos were obtained using the same cloud-based video recording solution and using the same endoscope manufacturer. Furthermore, while the algorithms performed well, our initial validation suggests that additional training is required for routine reliable use.

Conclusions

In summary, we describe the development and initial validation of the AI-CQ, an interactive AI-based tool to measure colonoscopy quality. While further improvements to the tool are planned, this interactive tool has the potential alter how we provide efficient and effective endoscopic training feedback and remediation.

Acknowledgement

This work was supported by the generous support of the Gordon and Betty Moore Foundation and the Northwestern Medicine Digestive Health Foundation.

Funding Information

Betty and Gordon Moore Foundation
Digestive Health Foundation

Conflict of Interest

Rajesh Keswani has served as a speaker and consultant for Boston Scientific and Medtronic. John Pandolfino has served as a speaker, a consultant, and an advisory board member for Ethicon, Endogastric Solutions, Medtronic, and Diversatek, and owns patent for FLIP Panometry. All other authors have no conflicts to disclose.

References

[1] Do A, Weinberg J, Kakkar A et al. Reliability of adenoma detection rate is based on procedural volume. Gastrointest Endosc 2013; 77: 376–380 doi:10.1016/j.gie.2012.10.023

[2] Pace D, Borgaonkar M, Evans B et al. Annual colonoscopy volume and maintenance of competency for surgeons. Surg Endosc 2017; 31: 2630–2635 doi:10.1007/s00464-016-5275-1

[3] Duloy A, Yadlapati RH, Benson M et al. Video-based assessments of colonoscopy inspection quality correlate with quality metrics and highlight areas for improvement. Clin Gastroenterol Hepatol 2019; 17: 691–700

[4] Duloy AM, Kaltenbach TR, Wood M et al. Colon polypectomy report card improves polypectomy competency: results of a prospective quality improvement study (with video). Gastrointest Endosc 2019; 89: 1212–1221

[5] Keswani RN, Byrd D, Garcia Vicente F et al. Amalgamation of cloud-based colonoscopy videos with patient-level metadata to facilitate large-scale machine learning. Endosc Int Open 2021; 9: E233–E238

[6] Shaukat A, Lichtenstein DR, Somers SC et al. Computer-aided detection improves adenomas per colonoscopy for screening and surveillance colonoscopy: A randomized trial. Gastroenterology 2022; 163: 732–741 doi:10.1053/j.gastro.2022.05.028

[7] Repici A, Badalamenti M, Maselli R et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 2020; 159: 512–520 e517 doi:10.1053/j.gastro.2020.04.062

[8] Glissen Brown JR, Mansour NM, Wang P et al. Deep learning computer-aided polyp detection reduces adenoma miss rate: A United States multi-center randomized tandem colonoscopy study (CADeT-CS Trial). Clin Gastroenterol Hepatol 2022; 20: 1499–1507 e1494

[9] Thakkar S, Carleton NM, Rao B et al. Use of artificial intelligence-based analytics from live colonoscopies to optimize the quality of the colonoscopy examination in real time: proof of concept. Gastroenterology 2020; 158: 1219–1221 e1212

[10] Barua I, Misawa M, Glissen Brown JR et al. Speedometer for withdrawal time monitoring during colonoscopy: a clinical implementation trial. Scand J Gastroenterol 2023; 58: 664–670

[11] Lux TJ, Sassmanshausen Z, Herold K et al. Assisted documentation as new focus for artificial intelligence in endoscopy: The precedent of reliable withdrawal time and image reporting. Endoscopy 2023; 55: 1118–1123

[12] Kaltenbach T, Patel SG, Nguyen-Vu T et al. Varied trainee competence in cold snare polypectomy – results of the COMPLETE randomized controlled trial. Am J Gastroenterol 2023; 118: 1880–1887 doi:10.14309/ajg.0000000000002368