Benchmarking definitions of false-positive alerts during computer-aided polyp detection in colonoscopy

Authors
Erik A. Holzwanger[1], Mohammad Bilal[2], Jeremy R. Glissen Brown[2], Shailendra Singh[3], Aymeric Becq[4], Kenneth Ernest-Suarez[5], Tyler M. Berzin[2]

Institutions
1 Division of Gastroenterology and Hepatology, Tufts Medical Center, Boston, Massachusetts, United States
2 Center for Advanced Endoscopy, Division of Gastroenterology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, United States
3 West Virginia University Health Sciences Center Charleston Division, Charleston, West Virginia, United States
4 Sorbonne Université, Centre d'Endoscopie Digestive, Hôpital Saint Antoine, APHP, Paris, France
5 Gastroenterology Department, Hospital México, University of Costa Rica, San Jose, Costa Rica

submitted 26.5.2020
accepted after revision 2.11.2020
published online 2.11.2020

Bibliography
Endoscopy 2021; 53: 937–940
DOI 10.1055/a-1302-2942
ISSN 0013-726X

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

Scan this QR-Code for the author commentary.

Corresponding author
Tyler M. Berzin, MD, Center for Advanced Endoscopy, Beth Israel Deaconess Medical Center, 330 Brookline Avenue, Boston, MA 02215, United States
tberzin@bidmc.harvard.edu

ABSTRACT

Background The occurrence of false-positive alerts is an important outcome measure in computer-aided colon polyp detection (CADe) studies. However, there is no consensus definition of a false positive in clinical trials evaluating CADe in colonoscopy. We aimed to study the diagnostic performance of CADe based on different threshold definitions for false-positive alerts.

Methods A previously validated CADe system was applied to screening/surveillance colonoscopy videos. Different thresholds for false-positive alerts were defined based on the time an alert box was continuously traced by the system. Primary outcomes were false-positive results and specificity using different threshold definitions of false positive.

Results 62 colonoscopies were analyzed. CADe specificity and accuracy were 93.2% and 97.8%, respectively, for a threshold definition of ≥0.5 seconds, 98.6% and 99.5% for a threshold definition of ≥1 second, and 99.8% and 99.9% for a threshold definition of ≥2 seconds.

Conclusion Our analysis demonstrated how different threshold definitions of false positive can impact the reported diagnostic performance of CADe for colon polyp detection.

Introduction

Over the past decade, there has been significant interest in applying artificial intelligence (AI) technologies to various areas in medicine. Machine learning, and a subset of machine learning termed "deep learning," are branches of AI centered on computer algorithms that can learn to perform a certain task, with performance that can improve over time with experience/training. An important application of AI and machine learning in colonoscopy is computer-aided detection (CADe) of colorectal polyps [1]. Several recent prospective trials have demonstrated that CADe may increase the adenoma detection rate [2–4].

An important outcome in most CADe colonoscopy studies is the number of false-positive alerts. A false positive is defined as

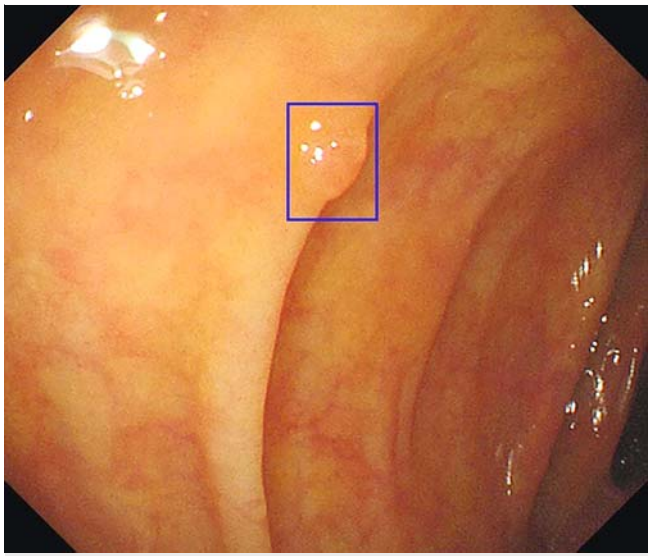
an area detected by the AI system that is not deemed to be a polyp by the endoscopist. Recent clinical studies on CADe have used a variety of definitions for false positive, ranging from time thresholds of > 1 or 2 seconds of an incorrect alert box, to vague definitions such as a nonpolyp area "continuously traced by the system," while some studies have not specified the definition of false positive at all [5 – 9]. There is currently no consensus definition of false positive for CADe. We aimed to study the diagnostic performance of CADe for colonoscopy based on different threshold definitions of false positive.

Methods

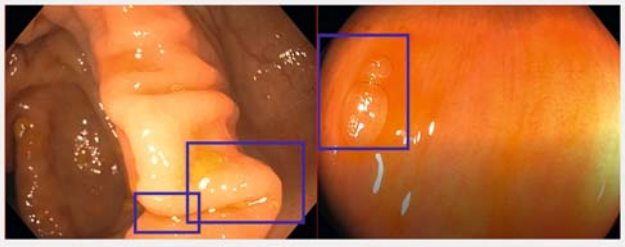
A previously validated deep learning CADe system (Shanghai Wision AI Co., Ltd., China) for polyp detection was applied to previously collected colonoscopy videos [7]. The CADe software uses a convolutional neural network based on SegNet architecture, which showed a high sensitivity and specificity for the detection of adenomas in previous validation data [3, 6,7]. The training and validation schemes for earlier versions of this algorithm have been detailed in previous studies, and the model has also been studied in prospective clinical trials [3, 8].

For the present study, colonoscopy videos were collected prospectively from September 2016 to March 2017 at a single endoscopy center in Costa Rica, with consecutive patients undergoing routine colonoscopy. Exclusion criteria included incomplete colonoscopy, history of colorectal cancer or active inflammatory bowel disease. We chose to use a time-based definition of false-positive alerts, as these have been used in previous clinical stage CADe studies and are likely to be more clinically relevant than other definitions and performance metrics used in preclinical development of CADe systems. Preclinical metrics such as precision recall were not included in the current analysis.

The AI-labeled videos were independently reviewed by a second gastroenterologist. When the AI system detects a polyp, a blue rectangular alert box appears on the screen around the area where a polyp is suspected. A true positive was defined as a polyp detected by AI for any length of time that was confirmed to be a polyp by the endoscopist (▶ **Fig. 1**). A false negative was defined as a polyp detected by the endoscopist but not detected by the AI system. A false positive was defined as an area detected by the AI system at any point that was not deemed to be a polyp by the endoscopist and second reviewer (▶ **Fig. 2**). The time duration for each false positive was recorded using a stopwatch. Per-polyp false positive was recorded rather than per-frame false positive. Frame-based definitions for false positives have been used during development and early testing of AI systems, but they are an unrealistic measure for clinical practice as seen in earlier AI studies [7]. Different thresholds for false-positive alerts were determined based on the time that a false-positive alert was continuously traced by the system. The different thresholds were: i) ≥ 0.5 seconds (Group 1), ii) ≥ 1 second (Group 2), and iii) ≥ 2 seconds (Group 3). False positives were categorized with respect to the actual endoscopic finding (mucosal fold, bubble, stool, or other). Withdrawal times and quality of bowel preparation using both the Boston Bowel Preparation Scale and the Aronchick Scale were collected.



▶ **Fig. 1** Example of a polyp detected by the artificial intelligence system (true positive).



▶ **Fig. 2** Examples of thick fold and stool debris (left) and a bubble (right) causing false-positive alerts by the computer-aided colon polyp detection system.

The primary outcome was number of false positives per colonoscopy, using the different false-positive thresholds. Secondary outcomes were specificity and accuracy of each false-positive group and comparison between different etiologies.

Statistical analysis was performed using STATA, version 14.0 (StataCorp, College Station, Texas, USA). Continuous variables were presented as mean and standard deviation (SD), whereas categorical variables were expressed as proportions and percentages. Continuous variables were compared using two-sample *t* test and categorical variables were compared using chi-squared test. Univariate logistic regression was performed to study factors associated with false positives. A two-sided *P* value of < 0.05 was considered statistically significant. Confidence intervals for specificity and accuracy were also calculated using "exact" Clopper – Pearson confidence intervals. The study was approved by the local Institutional Review Board.

▶ **Table 1** Patient and procedural characteristics

Demographics	
▪ Mean age, years	63.2
Sex, n (%)	
▪ Male	24(38.7)
▪ Female	38 (61.3)
Procedure indications, n (%)	
▪ Colon cancer screening	48 (77.4)
▪ Surveillance colonoscopy	4 (6.5)
▪ Abdominal pain	5 (8.1)
▪ Diarrhea	2 (3.2)
▪ Other	3 (4.8)
Bowel preparation	
▪ BBPS, mean (SD)	8.3 (0.7)
Aronchick scale, n (%)	
▪ Excellent	28 (45.2)
▪ Good	21 (33.9)
▪ Fair	12 (19.3)
▪ Inadequate	0
▪ Poor	1 (1.6)

SD, standard deviation; BBPS, Boston Bowel Preparation Scale.

Results

A total of 62 colonoscopy videos were included in the study. Patient and procedural characteristics are shown in ▶ **Table 1**. At least one polyp was detected in 42 colonoscopies by the endoscopists. A total of 95 polyps (true positives) were detected. There were no false negatives (none of the polyps detected by the endoscopist were missed by the AI system).

A total of 1635 false positives were seen: 91.8 % were folds, 5.6 % were bubbles, and 2.5 % were defined as stool or other. The number of false positives varied in different groups based on the respective time threshold definitions of false positive. A total of 1498 false positives were seen only "instantaneously" (< 0.5 seconds) and did not meet our time threshold definitions. There were 111 false positives in Group 1 (≥ 0.5 seconds), 23 in Group 2 (≥ 1 second), and 3 in Group 3 (≥ 2 seconds) (▶ **Table 2**).

The CADe system detected all actual polyps in all groups (true positives) (▶ **Fig. 1**), and missed none of the 95 polyps detected by the endoscopists (false negatives). The specificity and accuracy varied based on the threshold time categories of false-positive alerts. With a false-positive threshold of > 0.5 seconds, specificity and accuracy values were 93.2 % and 97.8 %. When the false-positive definition was changed to ≥ 2 seconds, specificity and accuracy were 99.8 % and 99.9 %, respectively (▶ **Table 2**).

Using the "instantaneous" threshold definition of false positive, the mean number of false positives was significantly higher in colonoscopies with fair or poor bowel preparation compared with excellent or good preparation (36.5 [SD 13.2] vs. 23.7 [SD 15.9]; $P < 0.01$). Given that the mean false-positive rate was 26.3 (using the "instantaneous" threshold), we then categorized colonoscopies as "high false-positive rate" (> 25 false positives) or "low false-positive rate" (≤ 25 false positives). High false-positive rate was associated with a fair or poor Aronchick bowel preparation score. We further analyzed false positives for colonoscopies with poor bowel preparation. The mean false-positive rate with fair or poor preparation in Group 1 was 1.9/colonoscopy (SD 1.1); however, there were no false positives in Group 3 (▶ **Table 2**). Longer withdrawal times were associated with higher false-positive rates.

Discussion

An ideal CADe system should have a high sensitivity for polyp detection, low rates of false-positive alerts, a low latency, and low cost per procedure [9, 10]. Understanding the behavior of CADe systems with respect to false positives is essential in com-

▶ **Table 2** Diagnostic performance of computer-aided detection using different thresholds for false-positive alerts.

	False-positive alert		
	≥ 0.5 seconds (Group 1)	**≥ 1 second (Group 2)**	**≥ 2 seconds (Group 3)**
Total false positives (62 colonoscopies)	111	23	3
False positives per colonoscopy, mean (SD)	1.8 (3.1)	0.4 (0.8)	0.05 (0.3)
False positives per colonoscopy with fair – poor bowel preparation, mean (SD)	1.9 (1.1)	0.2 (0.4)	0
Specificity, % (95 %CI)	93.2 (91.9 – 94.4)	98.6 (97.9 – 99.1)	99.8 (99.5 – 99.9)
Accuracy, % (95 %CI)	97.8 (97.0 – 98.4)	99.5 (99.1 – 99.8)	99.9 (99.7 – 100.0)

SD, standard deviation; CI, confidence interval.

paring the performance of these technologies. Our current analysis revealed how different threshold definitions of false positive can dramatically impact reported false-positive results and influence the perceived diagnostic performance of CADe for polyp detection. In Group 1, there were 111 false-positive alerts, whereas in Group 3 (≥2 seconds), only 3 false-positive alerts were noted. These results had a significant impact on the specificity and thus the accuracy of the CADe system.

Using different benchmarks for false-positive alerts can lead to difficulty in comparing the performance of different CADe systems, and many studies do not explicitly define false positives at all. We suggest that a consensus benchmark for defining false positives is needed to standardize the interpretation of data for CADe in colonoscopy. We propose that a ≥2-second threshold may be most appropriate and practical for defining false positives in CADe for colon polyp detection. A 2-second definition allows time for bubbles/debris to be irrigated away and for folds to flatten with insufflation, both of which are standard techniques during high quality colonoscopy; after 2 seconds, the few alerts boxes that remain must be carefully defined as false positives. A standardized approach to false-positive definitions will not only help in determining the true accuracy and specificity of CADe systems, but it will allow for more accurate comparison between different CADe systems.

Our study also revealed that poor bowel preparation was an independent factor for increased number of false-positive alerts, as seen in previous studies revealing many false-positive alerts consisting of stool/bubbles [8]. This has important clinical relevance outside of the clinical trial setting, as suboptimal bowel preparation is common and endoscopists may still have to make reasonable attempts at polyp detection.

Our study has several important limitations. Although we utilized prospectively collected colonoscopy videos, the AI analysis was performed after the procedure. Software design in computer-aided polyp detection is rapidly improving, and other or newer iterations of the CADe systems may perform differently and are already being incorporated into subsequent studies. Additionally, while time-based thresholds for false-positive alerts are more clinically relevant than the frame-based or per-image-frame false-positive definitions often used in the preclinical development of CADe systems, these time-based false-positive definitions can be influenced by endoscopist technique, including speed of withdrawal. Our proposed ≥2-second false-positive threshold requires methodical investigation of any areas where alert boxes are seen. In addition, as this was a post hoc analysis of already collected videos, we did not evaluate the impact of time-based definitions for AI alerts on true positives; thus, a sensitivity calculation could not be accurately performed. Finally, a question remains regarding whether alerts that appear in response to an area with bubble or stool that disappear after appropriate cleaning should be considered as false positives or not, and previous approaches to this question have varied. We feel that an alert box that is present for >2 seconds (after irrigation and insufflation) is relevant for practicing gastroenterologists and should be reported as a false positive.

To our knowledge, this is the first study to evaluate the impact that various false-positive time thresholds can have on the perceived diagnostic performance of a computer-aided polyp detection system. As the field of CADe continues to progress rapidly, establishing consensus definitions for CADe performance parameters is of significant value. We suggest that a false-positive threshold of ≥2 seconds is a clinically reasonable and practical starting point. Future studies using other CADe algorithms are needed to confirm that our suggested threshold of ≥2 seconds is applicable across different CADe systems.

Competing interests

Tyler Berzin is a consultant for Wision AI, Fujifilm, Boston Scientific, Magentiq Eye, and Medtronic. All other authors declare that they have no conflicts of interest.

References

[1] Mori Y, Kudo SE, Misawa M et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. Ann Intern Med 2018; 169: 357–366

[2] Urban G, Tripathi P, Alkayali T et al. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 2018; 155: 1069–1078

[3] Wang P, Berzin TM, Glissen Brown JR et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomized controlled study. Gut 2019; 68: 1813–1819

[4] 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

[5] Misawa M, Kudo SE, Mori Y et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology 2018; 154: 2027–2029

[6] Wang P, Liu X, Berzin TM et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol Hepatol 2020; 5: 343–351

[7] Wang Z, Liang Z, Li L et al. Reduction of false positives by internal features for polyp detection in CT-based virtual colonoscopy. Med Phys 2005; 32: 3602–3616

[8] Wang P, Xiao X, Glissen Brown JR et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biomed Eng 2018; 2: 741–748

[9] Vinsard DG, Mori Y, Misawa M et al. Quality assurance of computer-aided detection and diagnosis in colonoscopy. Gastrointest Endosc 2019; 90: 55–63

[10] Alagappan M, Brown JRG, Mori Y et al. Artificial intelligence in gastrointestinal endoscopy: the future is almost here. World J Gastrointest Endosc 2018; 10: 239–249