EDITORIAL

Meta-analyses of machine learning in endoscopy: stacking apples and oranges

The endoscopic literature is currently overwhelmed by publications on machine learning: "the use of mathematical algorithms (often nicknamed as artificial intelligence) for capturing structure in endoscopic images."[1]

Machine learning in endoscopy holds several promises, of which improving the detection of early neoplastic lesions (ie, computer-aided detection [CAD]) is most appealing: most GI cancers are diagnosed at an advanced stage, which carries a poor prognosis and/or requires invasive treatment. When they are detected at an early stage, the chances of cure after minimally invasive treatment are excellent. Early neoplastic lesions, however, are asymptomatic and can be detected during endoscopy only as a coincidental finding or as part of endoscopic screening or surveillance (eg, Barrett's esophagus). Because early neoplasia may not be easily recognized by endoscopists, given their subtle endoscopic appearance and low prevalence in the general population, improving detection by CAD algorithms may improve patient outcomes. Building such a CAD algorithm starts by a training phase on available endoscopic images of early neoplasia to identify relevant features and their relative weights within the training set. These features and weights can then be used to have the algorithm make predictions on a new, unseen set of endoscopy images (external validation) before being tested during real-time endoscopy.

Most current machine learning publications report outcomes of preclinical studies that are often driven by the simple availability of a set of endoscopic images or videos and associated with important methodologic pitfalls.[1] The most important 3 pitfalls are these: (1) selection bias: only high-quality detailed images of neoplasia, obtained by expert endoscopists, are used to train and validate the algorithm; (2) overfitting: the algorithm originates from multiple train-and-test-iterations within the same dataset to such an extent that it perfectly describes the training set but does not allow accurate predictions on new, unseen images; and (3) lack of independent external validation (ie, the algorithm is not tested on a new, unseen set of images).

In this issue of *Gastrointestinal Endoscopy*, Bang et al[2] present a meta-analysis of all available preclinical studies of CAD systems for early esophageal neoplasia. The authors justify the use of a meta-analysis to overcome the limitation of the small sample sizes of independent studies and to provide a more accurate estimation of CAD performance. In this meta-analysis, the authors attempt to group the performance of preclinical CAD systems for the endoscopic detection of early neoplasia in overview, the characterization of lesions (neoplastic vs nonneoplastic) upon detailed inspection, and the estimation of invasion depth of Barrett's neoplasia, squamous cell carcinoma, and "other" esophageal neoplasia. The included CAD systems were developed by the use of high-definition white-light endoscopy, narrow-band imaging, or blue-light imaging, in overview or magnification. It is thereby one of the few systematic reviews to report the pooled diagnostic accuracy of upper GI CAD systems.[3-5]

Our practical advice here is to be careful when reading preclinical CAD articles: basically, don't even bother to read the article if the presented algorithm is not explicitly tested on an independent external dataset, and even then, be aware of the various forms of both visible and invisible bias.

For the results of studies to be grouped in a meta-analysis, the studies need to be homogenous for their design and outcome parameters, and only studies with adequate quality should be included. The heterogeneity of the studies selected for inclusion by Bang et al[2] appears to be large, given the types of pathologic conditions (Barrett's and squamous neoplasia), imaging technology (white-light endoscopy in overview vs detailed optical chromoscopic images), and envisioned applications of the various CAD systems (detection, characterization, and invasion depth estimation). Surprisingly, the authors report low or even nonexistent statistical heterogeneity. This seems contradictory, given the clear clinical differences between the included studies. This can partly be explained by the rationale that commonly used statistical methods for heterogeneity may be less applicable for preclinical CAD studies. An example is the observation of the shape of the summary receiver operating characteristic (SROC) curve, one of the methods used by

0016-5107/$36.00
https://doi.org/10.1016/j.gie.2021.01.010

the authors to test heterogeneity. In general, reported diagnostic accuracy in preclinical CAD studies is high. When results are pooled in SROC curves, these may be symmetric, notwithstanding heterogeneity in experimental setups as mentioned above. In addition, it should be noted that the I^2 estimates of specificity and sensitivity in this meta-analysis do in fact suggest statistical heterogeneity.

Another important element in any meta-analysis is the quality assessment of included studies, for which the authors use the QUADAS-2 tool, a standardized tool to evaluate risk of bias in diagnostic accuracy studies. Although commonly accepted, this tool is less suited for the quality assessment of CAD studies, where the methodology and experimental setup of preclinical studies often introduce a variety of types of bias.[1] When the QUADAS-2 tool was used, nearly all included studies were scored as having a low risk of bias, whereas in fact all of these studies (including those reported by our group) are significantly hampered by, for example, selection bias of the type and quality of images used for training the algorithms. The issue of quality assessment for CAD studies has led others to adapt the QUADAS-2 tool, identifying specific bias domains for diagnostic studies in machine learning.[3]

Given these limitations with regard to heterogeneity and quality assessment and with the many pitfalls in designing and interpreting preclinical CAD studies, the additive value of a meta-analysis to pool CAD performance is therefore debatable. As mentioned by Bang et al,[2] there is, however, a need for a more reliable estimation of the value of CAD systems in endoscopy.

What is the best way forward to achieve this goal? The answer is twofold: we will need to (1) more critically appraise the plethora of preclinical CAD studies and (2) insist on in-vivo clinical studies with appropriate designs and clinically relevant endpoints.

Preclinical CAD studies should be more critically appraised. Authors should adhere to the basic minimum requirements for methodologic and experimental setup.[1]

Recently, machine learning–specific extensions of the CONSORT guideline and the SPIRIT guideline have been proposed.[6,7] These guidelines, however, propose general recommendations for a broad variety of applications of artificial intelligence in gastroenterology (eg, including nonendoscopic prediction models), thereby limiting their applicability for endoscopic studies. To provide explicit guidance for authors, reviewers, and editors of peer-reviewed journals, endoscopy-specific guidelines are warranted. Also, to objectively evaluate performance and enable direct comparison among CAD systems, large endoscopic datasets should be collected by endoscopy societies, on which CAD systems can be tested for performance thresholds to clinical implementation.[1] Ideally, for each specific CAD application (eg, polyp classification, primary detection of gastric neoplasia) a separate test set should be available. These datasets should meet general quality requirements, such as sufficient sample size and heterogeneity in terms of image quality and center of origin. Needless to say, creating and updating these datasets and safeguarding them from being used only for testing performance without prior training will pose major challenges.

There is an increasing need for well-designed randomized controlled trials that evaluate the additive value of CAD systems in daily practice. Recently, several articles were published describing randomized controlled trials for polyp detection in the colon.[8] These studies all showed an increase in the adenoma detection rate, but none of them showed a significant increase in the detection of advanced adenomas. Which of the following are relevant endpoints for such a randomized controlled trial: any additional polyp detected with CAD, any additional adenoma, any additional advanced adenoma, or only additionally detected advanced adenomas in patients who would otherwise be pushed for a long-term surveillance interval? Not everything we miss has relevant clinical consequences.

The argument can be made that the clinical impact of such detection algorithms may be inferior to that of quality assurance algorithms. Such algorithms focus on optimizing the quality of the endoscopic procedure by interacting with the endoscopist, eg, by providing feedback on withdrawal speed, percentage of mucosa visualized, or bowel preparation. Endoscopists might hate this "big-brother-is-watching-you-during-endoscopy" algorithm; yet, its impact will likely be larger than a CAD system showing them diminutive polyps for which overlooking may even be the best approach for the patient.

Machine learning in endoscopy holds many promises. The most appealing is to eliminate the human factor with regard to "recognizing what can be seen" and eliminating variability in mindset and concentration. Now that the feasibility of its use is apparent, it is time that we start appraising scientific output more critically and focus on how CAD will affect patient outcomes in daily endoscopic practice. Our practical advice here is to be careful when reading preclinical CAD articles: basically, don't even bother to read the article if the presented algorithm is not explicitly tested on an independent external dataset, and even then, be aware of the various forms of both visible and invisible bias. During the next years we will see more "proof of pudding" studies in which algorithms are tested in large-scale randomized clinical studies. Here the main focus during critical appraisal should be on the relevance of the reported endpoint: not every additional lesion that is detected bears clinical relevance. Otherwise, we run the risk that machine learning will mainly drive the already significant overconsumption in endoscopy instead of truly improving the quality of care of our patients.[9]

DISCLOSURE

All authors disclosed no financial relationships.

Jeroen de Groof, MD, PhD
Department of Gastroenterology and Hepatology
Amsterdam UMC
Amsterdam, The Netherlands
Giulio Antonelli, MD, PhD
Gastroenterology Unit
Nuovo Regina Margherita Hospital
Department of Translational and Precision Medicine
Sapienza University of Rome
Maria J. Dinis-Ribeiro, MD, PhD
Department of Translational and Precision Medicine
Instituto Portugues de Oncologia do Porto
Portugal
Jacques J. Bergman, MD, PhD
Department of Gastroenterology and Hepatology
Amsterdam UMC
Amsterdam, The Netherlands

Abbreviation: CAD, computer-aided detection.

REFERENCES

1. van der Sommen F, de Groof J, Struyvenberg M, et al. Machine learning in GI endoscopy: practical guidance in how to interpret a novel field. Gut 2020;69:2035-45.
2. Bang CS, Lee JJ, Baik GH. Computer-aided diagnosis of esophageal cancer and neoplasms in endoscopic images: a systematic review and meta-analysis of diagnostic test accuracy. Gastrointest Endosc 2021;93:1006-15.e13.
3. Arribas J, Antonelli G, Frazzoni L, et al. Standalone performance of artificial intelligence for upper GI neoplasia: a meta-analysis. Gut. Epub 2020 Oct 30.
4. Jin P, Ji X, Kang W, et al. Artificial intelligence in gastric cancer: a systematic review. J Cancer Res Clin Oncol 2020;146:2339-50.
5. Lui TKL, Tsui VWM, Leung WK. Accuracy of artificial intelligence-assisted detection of upper GI lesions: a systematic review and meta-analysis. Gastrointest Endosc 2020;92:821-30.e9.
6. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020;370:m3164.
7. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health 2020;2:e549-60.
8. Hassan C, Spadaccini M, Iannone A, et al. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. Gastrointest Endosc 2021;93:77-85.e6.
9. Shaheen NJ, Fennerty MB, Bergman JJ. Less is more: a minimalist approach to endoscopy. Gastroenterology 2018;154:1993-2003.