EDITORIAL

Artificial intelligence for dysplasia grading in Barrett's esophagus: hematoxylin and eosin is here to stay

Artificial intelligence (AI) is defined by the *Oxford English Dictionary* as "the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages." Although still in its infancy, AI has already globally transformed several aspects of 21st century technology, influencing aviation, advertising, law enforcement, and warfare.[1-3] It is only natural that AI will also transform medicine and improve patient care.

Because AI is especially adept at image analysis, pathology is one field of medicine that seems extremely well suited to reap the benefits of this technology. AI learns to perform a particular task by first gaining exposure to large numbers of representative examples of that task (training sets). The composition of the training set is therefore a crucial determinant of how well AI will perform the task at hand. Training sets consisting of relatively simple gross endoscopic images have already been used to create AI algorithms that assist gastroenterologists in sampling lesions and avoiding blind spots. Training sets composed of predominantly black-and-white images for AI-guided radiology interpretation are another success story of applying the power of AI image analysis to medicine. A major reason for these successes is the training sets, the most successful of which leverage large numbers of relatively straightforward images with manageable variance across the same diagnostic entity and comparable image quality between different institutions.

By contrast, the hematoxylin and eosin (H&E) stained glass slides required to create histology images for effective pathology training sets are complex. In particular, there is often substantial morphologic variance across a single diagnostic entity and widely divergent processing quality between institutions. Glass slides must also be electronically scanned and converted into digital image formats. Processing artifacts, the large number of diagnostic entities, and the (often) small numbers of representative cases introduce additional barriers to constructing effective AI training sets. Because of these and other issues, it has proved more difficult to apply AI image analysis to pathology than to other fields of medicine. A major breakthrough toward this end was the development of high-quality scanners that rapidly convert glass H&E slides into digital image files, which first became available between 2005 and 2007. However, even this advance introduced a new set of practical barriers. The scanners themselves remain very expensive, ample physical space must be designated to house the scanners, and the scanned images require thousands of terabytes of storage space.[4] There are also unique technical issues. For example, slides can be inadvertently scanned out of focus, and the presence of air bubbles or other artifacts on the slides can also interfere with image quality. Despite these problems, some pathology departments in the United States have successfully introduced routine digitization of glass slides, and there are proof-of-principle examples of using these images to train AI for certain diagnostic pathology applications.[4]

> Because AI is especially adept at image analysis, pathology is one field of medicine that seems extremely well suited to reap the benefits of this technology.

In this issue of *Gastrointestinal Endoscopy*, Faghani et al[5] report their exploration of the feasibility of using AI to detect and grade Barrett's esophagus–associated dysplasia. Other investigators have reported favorable results applying AI for this purpose,[6] but the current study is a significant improvement because of superior training sets consisting of an unusually large number of cases (slides) that were collected and scanned at the Mayo Clinic under the supervision of 2 expert GI pathologists.

Patients with a primary diagnosis of unablated Barrett's dysplasia were included in the study. Slides were retrospectively re-examined by the study pathologists. If there was agreement with the initial interpretation, the slide was used for the training set. If there was no agreement, the slide was excluded. Training set slides were then digitized into whole slide images by use of a high-resolution slide scanner and digitally annotated ("supervised") by the study pathologists, who marked the highest grade of dysplasia on each slide (as the "ground truth"). This yielded a set of 620 slides for scanning, although 78 were eliminated because of processing artifacts

0016-5107/$36.00
https://doi.org/10.1016/j.gie.2022.08.013

(pale or faded stains). Ultimately, a training set of 368 whole imaged slides representing the range of dysplasia categories in Barrett's esophagus patients (negative for dysplasia, low-grade dysplasia, and high-grade dysplasia) was constructed to train the AI program. The post-training validation set consisted of 104 additional slides to fine-tune the program, and a final test set of 70 new slides (not previously seen by the program) was used to formally evaluate performance. The authors used a YOLO (You Only Look Once) model for whole slide image analysis combined with a classifier to achieve excellent AI agreement with the 2 expert GI pathologists participating in the study.

The work by Faghani et al[5] is important and instructive. It specifically highlights the potential for AI-guided rapid screening of digitized slides to speed up-front detection and improve observer consistency for a specific, user-defined diagnostic spectrum (Barrett's dysplasias). Beyond the results, it is instructive to consider the precise details of how the AI algorithm was trained. In this case, achieving the exceptional performance required an unusually large training set that focused on a very narrow range of diagnostic possibilities using slides and images of uniform quality collated from a single institution. Such rigid requirements of the training set embody the difficulties that lie ahead for the successful implementation of AI technology into routine pathology practice on global, national, regional, or even local scales. Thus, although the study of Faghani et al[5] and others clearly indicate that AI technology can provide powerful new tools that improve the speed, efficiency, and consistency of human pathologists, they conversely indicate that AI in its current form is probably too inflexible to enable full automation of the final diagnosis because of the intrinsic limitations imposed by the rigorous training set requirements. To illustrate these points, we will highlight a handful of concrete situations for which sole reliance on AI for a final diagnosis of Barrett's dysplasia could become problematic.

First, it is unlikely that an algorithm trained to extract features of Barrett's dysplasia would also "flag" other incidental entities that inevitably appear on slides in routine practice, such as GI stromal tumors, lymphomas, or subtle signet ring cell adenocarcinomas. Achieving this ability will require especially sophisticated AI research, such as merging AI algorithms trained for different diagnostic categories into a single all-encompassing program or running each individual algorithm on the scanned images in succession. It seems unlikely that training sets will ever truly represent such scenarios in pathology, especially given the wide range of combinations and limited numbers of cases.

Second, it is unlikely that even the best trained AI algorithm would recognize uncommon Barrett's-related entities (such as dysplasia with gastric-type differentiation), again because of underrepresentation in the all-important training sets. This scenario was observed when AI was successfully applied for diagnosis of prostate adenocarcinoma.[7] Despite a training set that included >1700 slides, the final algorithm failed to detect several variant patterns of adenocarcinoma that were detected on tandem blind review by human experts. Similarly, an analogous system designed to detect carcinoma cells in pleural fluid was outperformed by senior cytopathologists, although AI did outperform their junior peers.[8]

Finally, common processing artifacts and differences in staining and sectioning quality that pose no diagnostic problems for human pathologists can baffle even the best-trained AI "pathologists." Unfortunately, such differences are often the rule rather than the exception between different institutions. This presents major disadvantages for constructing "universal" AI training sets for pathology. Lack of uniform slide processing may also preclude combining images from different institutions. This will place severe restrictions on constructing test sets large enough to effectively train AI for all but the most common or straightforward diagnostic entities. Even though the impressive study of Faghani et al[5] was conducted at a single, deeply experienced institution, a sizeable number of slides still had to be eliminated from the training set because of histology processing artifacts (faded stains).

Regardless of these and other issues, AI will inevitably become a valuable or even indispensable tool for routine pathology practice. At the very least, this technology will streamline practice efficiency by screening cases and prioritizing case review, improve diagnostic consistency in gray areas prone to observer variation, and potentially elevate generalist pathologists to the level of specialist experts by flagging exquisite details that would not have been appreciated otherwise (owing to lack of subspecialty training). AI will also be convenient. Analyses that are tedious and uninteresting for human pathologists, such as immunohistochemical scoring (for example, PD-L1) or screening lymph nodes for metastatic disease,[9,10] will become partially if not exclusively automated. These exciting possibilities will be achieved because AI is especially adept at image analysis, which is the foundation of diagnostic pathology. It is somewhat ironic that after all the molecular advances made over the previous decades, the humble H&E-stained glass microscope slide will remain a gold standard for diagnosing human disease.

DISCLOSURE

Dr McDonald is the recipient of funding from National Institutes of Health grant R01 CA222594. The other author disclosed no financial relationships.

Oliver G. McDonald, MD, PhD
Elizabeth A. Montgomery, MD
Department of Pathology and Laboratory Medicine
University of Miami Miller School of Medicine
Miami, Florida, USA

Abbreviations: AI, artificial intelligence; H&E, hematoxylin and eosin.

REFERENCES

1. Avdelidis NP, Tsourdos A, Lafiosca P, et al. Defects recognition algorithm development from visual UAV inspections. Sensors (Basel) 2022;22.
2. Mohan PV, Dixit S, Gyaneshwar A, et al. Leveraging Computational Intelligence Techniques for Defensive Deception: A Review, Recent Advances, Open Problems and Future Directions. Sensors (Basel) 2022;22:4682.
3. Sun H, Zafar MZ, Hasan N. Employing natural language processing as artificial intelligence for analyzing consumer opinion toward advertisement. Front Psychol 2022;13:856663.
4. Hanna MG, Ardon O, Reuter VE, et al. Integrating digital pathology into clinical practice. Mod Pathol 2022;35:152-64.
5. Faghani S, Codipilly DC, Vogelsang D. Development of a deep learning model for the histological diagnosis of dysplasia in Barrett's esophagus. Gastrointest Endosc 2022;96:918-25.e3.
6. Sali R, Moradinasab N, Guleria S, et al. Deep learning for whole-slide tissue histopathology classification: a comparative study in the identification of dysplastic and non-dysplastic Barrett's esophagus. J Pers Med 2020;10:141.
7. Perincheri S, Levi AW, Celli R, et al. An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. Mod Pathol 2021;34:1588-95.
8. Xie X, Fu CC, Lv L, et al. Deep convolutional neural network-based classification of cancer cells on cytological pleural effusion images. Mod Pathol 2022;35:609-14.
9. Huang SC, Chen CC, Lan J, et al. Deep neural network trained on gigapixel images improves lymph node metastasis detection in clinical settings. Nat Commun 2022;13:3347.
10. Wu J, Liu C, Liu X, et al. Artificial intelligence-assisted system for precision diagnosis of PD-L1 expression in non-small cell lung cancer. Mod Pathol 2022;35:403-11.