NARRATIVE REVIEWS

Charles J. Kahi, Section Editor

The Evolving Role of Artificial Intelligence in Gastrointestinal Histopathology: An Update

D. Chamil Codipilly,[1] Shahriar Faghani,[2] Catherine Hagan,[3] Jason Lewis,[4] Bradley J. Erickson,[2] and Prasad G. Iyer[1]

[1]Barrett's Esophagus Unit, Division of Gastroenterology and Hepatology, Mayo Clinic Rochester, Rochester, Minnesota; [2]Mayo Artificial Intelligence Laboratory, Department of Radiology, Mayo Clinic, Rochester, Minnesota; [3]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota; and [4]Department of Pathology, Mayo Clinic, Jacksonville, Florida

Significant advances in artificial intelligence (AI) over the past decade potentially may lead to dramatic effects on clinical practice. Digitized histology represents an area ripe for AI implementation. We describe several current needs within the world of gastrointestinal histopathology, and outline, using currently studied models, how AI potentially can address them. We also highlight pitfalls as AI makes inroads into clinical practice.

Keywords: Artificial Intelligence; Deep Learning; Digital Histopathology.

Although the inception of artificial intelligence (AI) began more than 50 years ago, there has been a dramatic expansion of AI research in the clinical realm over the past decade.[1] As researchers and institutions, both public and private, have created increasingly complex AI models, clinicians have been promised a paradigm shift in the management of patients, both at the population as well as individual levels. As of this writing, AI applications in general clinical use include virtual chatbots that can parse through patient symptoms and recommend whether further (and what specific) evaluation is required,[2,3] wearable technology that can assess for the presence of atrial fibrillation,[4] models that can assist clinicians with real-time colon polyp identification,[5] and algorithms that can aggregate hospital-based outcomes enabling comparisons with other institutions.[6] Notably, the vast majority of Food and Drug Administration (FDA)-cleared devices are in radiology, likely because of the large archives of digital data that are critical for the development of AI models.[7] It is abundantly clear that given its clinical and procedural emphasis, gastroenterology is a field that is ripe for innovation from an AI standpoint. Indeed, a simple PubMed search for *artificial intelligence gastroenterology* has seen a distinct increase in annual publications, from 13 in 2013 to more than 600 in 2022.

The clinical adoption of high-throughput digitization of pathology slides has accelerated in the past 5 years, particularly after the FDA approved marketing of a digitized whole slide imaging platform.[8] Given the critical role histology plays in the management of an array of gastrointestinal disorders, AI potentially will impact diagnosis and therapy in this field.

In this review we provide an overview of AI applications for gastrointestinal histopathology near adoption or currently in clinical use, highlighting those that will have a significant impact on clinical care. It is imperative that practicing physicians understand the implications of AI, including potential contributions, pitfalls, and impact to their profession.

What Is Artificial Intelligence?

The term *artificial intelligence* was coined in the 1950s initially as a framework of *if, then* statements, with the goal of simulating critical thinking.[9] With advances in algorithmic development, AI models in the modern area can complete complex tasks with accuracy and efficacy on par with human beings, a process that has been expedited by the creation of machine learning (ML) and deep learning models.

Analysis of histopathology images using ML can be considered as a subset of computer vision. However, it is important to recognize that other types of data can be leveraged for image-oriented tasks, including unstructured text notes of an electronic medical record that can be processed via natural language processing models,[10] as well as tabular data, defined as information arranged in rows and columns, such as laboratory values. This non-imaging information can be merged with histopathology inputs when training a model.

Table 1 presents an overview of the various tasks that can be performed in histopathology using

Abbreviations used in this paper: **AI, artificial intelligence; AUC, area under the curve; BE, Barrett's esophagus; CNN, convolutional neural network; CRC, colorectal cancer; EAC, esophageal adenocarcinoma; EoE, eosinophilic esophagitis; FDA, Food and Drug Administration; GI, gastrointestinal; HCC, hepatocellular carcinoma; ML, machine learning; NASH, nonalcoholic steatohepatitis; ROSE, rapid on-site evaluation; UC, ulcerative colitis; WATS-3D, wide-area transepithelial sampling with 3-dimensional analysis.**

Most current article

1542-3565/$36.00
https://doi.org/10.1016/j.cgh.2023.11.044

Table 1. Summary of Different Tasks and Corresponding Deep Learning–Based Computer Vision Approaches

Task	Description
Classification task	Assigning a class label to input examples (eg, deciding whether a WSI containing a colorectal polyp specimen is benign or malignant)
Segmentation task	Dividing different regions of an image by accurate delineation (eg, tumoral area on a pancreatic carcinoma WSI)
Object-detection task	Creating bounding boxes around the object of interest (eg, drawing bounding boxes around lymphocytes in gastric crypts)
Regression task	Predicting a continuous numeric value (eg, predicting free-recurrence survival from colorectal carcinoma WSI)
Image-generation task	Normalizing or synthesizing images that do not exist (eg, generating WSI with the staining style of Institute A using WSIs from Institute B)

WSI, whole slide image.

computer vision. Each of these tasks can be trained in a supervised manner, that is, if human annotations are provided to teach the model. The performance of the model will depend on the quality of annotations that were provided, and therefore annotation agreement often may be required among experts within the field of interest before including such data in model training. However, there also are models built in an unsupervised manner, in which human annotations are not included during training. This can be used in various tasks in which the objective is to discern different categories or clusters of information based on imaging features. Oftentimes, algorithms initially trained in an unsupervised manner form the basis of foundational models, which are finetuned using a supervised approach to improve performance.

Once developed, models then are validated, involving cross-fold validation to assess the model robustness, then evaluating performance on a hold-out internal test set, and, finally, assessing performance on an external test set (Figure 1). External validation of models is important to establish generalizability across different

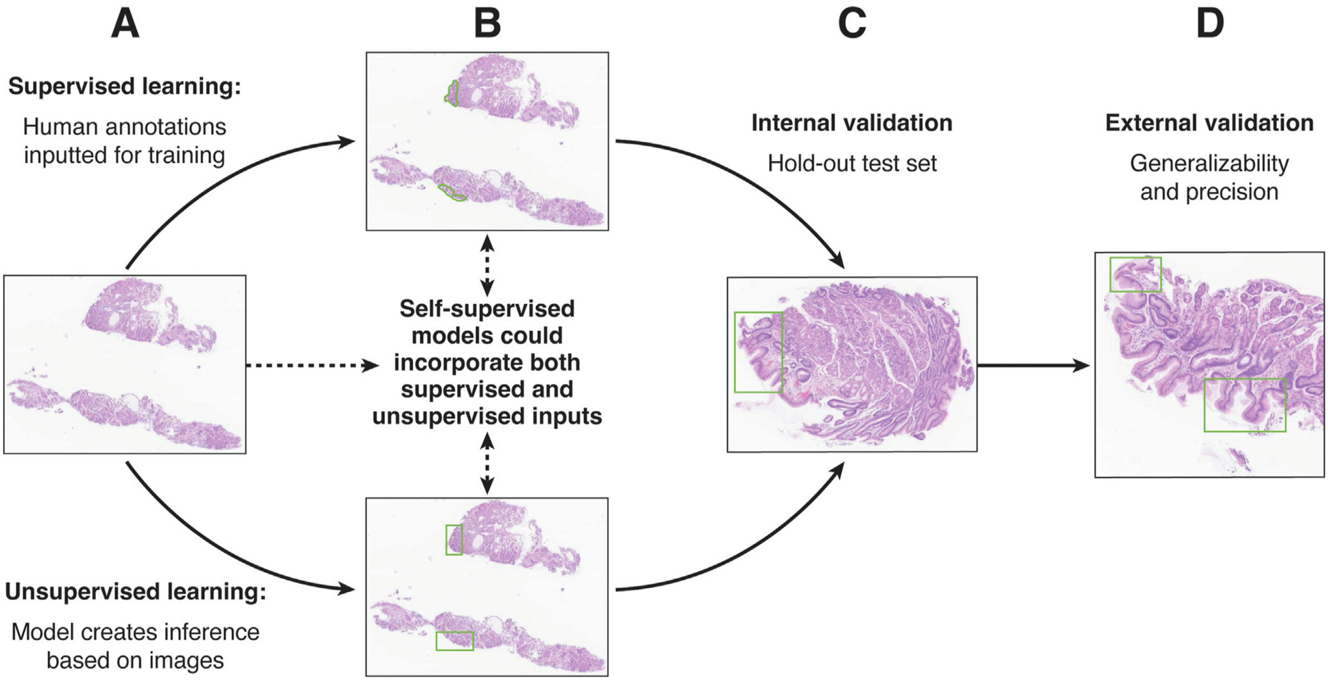
A
Supervised learning:
Human annotations inputted for training
Unsupervised learning:
Model creates inference based on images
B
Self-supervised models could incorporate both supervised and unsupervised inputs
C
Internal validation
Hold-out test set
D
External validation
Generalizability and precision

Figure 1. Schema for the development of an artificial intelligence model in digital histopathology. (*A*) Deep learning models may be trained in a supervised (annotations provided), unsupervised (no annotations provided), or combination manner for model training. (*B*) Unsupervised models could be adjusted based on inputs provided by researchers. (*C*) Model performance first is validated internally to assess performance characteristics, and (*D*) then is applied to an external validation set to assess generalizability and precision.

staining, scanning, and biopsy techniques, and tissue preparation types.

Opportunities for Improvement and Potential of Digital Pathology

The digitization of pathology enables facile implementation of AI into clinical workflows. There are a number of use cases in which AI involvement can benefit clinical care. Given the critical shortage of physicians in the United States,[11] the adoption of AI in clinical pathology workflow may allow for high-throughput slide assessment and initial impressions, which then can be reviewed by pathologists who can synthesize the slide findings within the clinical context. Furthermore, this may enable remote pathologist review (which is particularly helpful in resource-scarce locations), ultimately enabling appropriate management, which can lead to improved patient care. In addition, in disease states in which review of slides involves tedious manual segmentation of histologic features, AI may expedite this process, again freeing up pathologists for more pertinent work.

It is exceedingly unlikely that AI models will replace physicians. There are intangible factors that physicians may pick up on that could affect patient care, such as showing empathy, which allows a patient to open up and share symptoms that perhaps initially were not shared willingly, or to accomplish procedures without the need for complex robotics/machinery (eg, consider what type of device would be needed to place a central line on a crashing patient). From a pathology perspective, currently trained models will not be able to generate hypotheses and lead investigations based on disease processes seen on slides. However, given the ability to mine large, multicategoric data sets, it is conceivable that AI models may identify associations that are not inherently obvious to clinicians, and when combined with histopathology, novel diagnostic and prognostic information may be gleaned, leading to personalized therapeutic options for patients.

Slide digitization may enable collaboration in both clinical care and research. Currently, slide review from external organizations is a laborious process involving identification of appropriate patients, calling of stored slides (which often are kept at remote locations in the case of historical slides used in research, associated with slide degradation, a known complication that occurs over time), and, finally, manual review, until slides are shipped to a partnering institution, exposing them to the prospect of getting lost or damaged. Digitization offers the ability for rapid review of slides globally while maintaining the quality of slides as they were created originally. Coupled with AI, this could significantly improve diagnostic capabilities and enhance research collaboration.

What Are the Problems That Artificial Intelligence Currently Is Solving in Digital Histopathology?

Improved Efficiency of Tissue Analysis for Dysplasia Assessment

Barrett's esophagus. Barrett's esophagus (BE) is the only known precursor lesion for the development of esophageal adenocarcinoma (EAC). Multiple gastroenterological societies have recommended endoscopic surveillance of patient with nondysplastic BE to identify dysplasia, which is amenable to endoscopic therapy.[12,13] During endoscopic surveillance, 4-quadrant biopsy specimens are taken in 1- to 2-cm increments.[14] However, this method, although rigorous, misses sampling of the vast majority of the BE segment.

Wide-area transepithelial sampling with 3-dimensional analysis (WATS-3D; CDx Diagnostics, Suffern, NY) overcomes this limitation by using a stiff brush to sample a significantly greater area of BE mucosa. Samples are processed enabling a 3D view of tissue in 0.5-μ sections, which consist of mainly cells and some stromal elements. Using a convolutional neural network (CNN) developed on a residual neural network model, regions suspicious for atypical and/or dysplastic epithelium are identified and presented to an interpreting pathologist for final diagnosis.[15] This process rapidly increases the speed of interpretation by highlighting areas most concerning for manual review. As such, this particular example of an AI application allows for more efficient analysis of a larger amount of tissue than would be obtained via traditional BE sampling, while adding only an average of 4.5 minutes to the total procedure time.[15]

A meta-analysis of 13 studies showed that the addition of WATS-3D to routine BE sampling yielded an increase in dysplasia diagnosis by 2.1%.[16] Given these results, WATS-3D, which is commercially available and has its own Current Procedural Terminology billing codes, is available for clinical use and a large, prospective multicenter trial is underway to determine the impact of this AI-augmented digital histology test on dysplasia detection in comparison with conventional Seattle protocol biopsies.

Eosinophilic esophagitis. Several disease states require careful manual review and quantification of subtle histology features on slides, a prime example being eosinophilic esophagitis (EoE) and the EoE histologic scoring system.[17] Although research suggests this system outperforms standard peak eosinophil count with regard to disease diagnosis and monitoring, adoption is slow given the significant manual effort required to assess not only eosinophil count, but other components of the EoE histologic scoring system including basal zone hyperplasia, eosinophilic abscesses, surface layering, and so forth.[17] AI models may expedite the process of slide review, enabling pathologists to perform more efficient review of straightforward cases, thus saving time for more focused review of difficult cases. An AI model for segmentation of many of the features mentioned earlier was trained on 40 EoE slides, and validated on 203 slides with interobserver variability on par with that of pathologists, when compared with an interpretation by an expert gastrointestinal (GI) pathologist.[18] This may result in meaningful clinical changes in management because this information may help guide therapeutic decisions.

Helicobacter pylori. *H pylori* has been identified as a class 1 carcinogen and is likely the most significant global risk factor for gastric carcinoma. Although noninvasive testing is available for detection of *H pylori*, this bacterium can be identified on H&E and immunohistochemical staining on gastric biopsy specimens. Review of whole slides to find these small bacteria, often no larger than 4 μm, is tedious. As such, a CNN model was trained on 477 slides to aid pathologists in the diagnosis of *H pylori* on gastric biopsy specimens. This then was tested on a subset of 87 slides in which further polymerase chain reaction or immunohistochemical staining was performed as the gold standard. This decision support model achieved a sensitivity of 100% and a specificity of 68.4% when compared with the gold standard, with an area under the curve (AUC) of 0.92. Results were not as robust, but still reasonable, for H&E-stained slides (AUC, 0.81), and shows a model could feasibly minimize time spent assessing for *H pylori*.[19]

Ulcerative colitis. It is well established that histologic remission in ulcerative colitis (UC) is associated with improved clinical outcomes and has been suggested as a therapeutic target for management of inflammatory bowel diseases.[20] To that end, UC patients undergo frequent endoscopic evaluations to determine disease activity in response to various treatments. However, given the large amount of tissue obtained during procedures and the inherent difficulty with grading disease severity, this is a time-consuming task, even for experienced pathologists.[21] A CNN model was developed and trained on 118 colonic biopsy specimens to predict histologic and clinical outcomes in patients with UC. In a test set of 375 patients, this model achieved a sensitivity and specificity exceeding 75% for histologic scoring, and using a histologic index remission score, the model was able to accurately risk-stratify patients (and exceeded that as proposed by 6 expert inflammatory bowel disease pathologists) who would experience a UC flare from those who remained in remission.[21,22]

Colorectal cancer screening. Colorectal cancer (CRC) remains a leading cause of morbidity and mortality worldwide, and CRC screening has been recommended in the general population starting at the age of 45 because early stage diagnosis dramatically improves survival.[23,24] In 2014, the National Colorectal Cancer Roundtable initiated the Every Community Strategic Plan with the goal of achieving 80% CRC screening rates in communities across the United States, which would equal roughly 15 million colonoscopies per annum.[25] This would greatly increase the need for pathologist review of polypectomy specimens.

To mitigate this increase in pathology workload, a graph neural network was trained on more than 5000 colon biopsy specimens with a wide range of pathologies annotated by a team of expert GI pathologists.[26] The model was trained to distinguish normal from abnormal tissue, with the clinical goal of saving valuable time spent viewing unremarkable slides. This model achieved an area under the receiver operating characteristic of 0.98 and an F1 score of 0.98 on internal validation sets, and area under the receiver operating characteristics/F1 scores exceeding 0.95 and 0.94, respectively, on external validation. This model would reduce the need to review approximately 54% of slides in the validation sets, and in practice could increase the efficiency of clinical flow. A separate study examined the effect of an augmented AI histopathology imaging system trained on 326 slides to aid pathologist diagnosis among tubular, villous, sessile serrated, and hyperplastic polyps.[27] This improved diagnostic accuracy from 73.9% to 80.8% ($P < .001$), but also lead to a longer slide evaluation time, adding an average of 4.8 seconds (95% CI, 3.0–6.5 s) compared with microscope assessment alone.[28] However, the authors of this study noted that slide evaluation time improved over the course of the study as pathologists became more familiar with the software.

Improved Diagnostic Accuracy

Barrett's esophagus dysplasia. The diagnosis of dysplasia in BE is difficult, and likely is impacted by the presence of nonspecific inflammatory and regenerative changes and subjective criteria for dysplasia. Consequently, agreement among pathologists for the diagnosis of dysplasia is suboptimal, with poor agreement (as measured by the κ statistic <0.5), a metric that has not improved in nearly 20 years.[29] This results in an abundance of overcalled dysplastic disease with important ramifications, potentially leading to unnecessary endoscopic procedures in patients without a high risk of neoplastic progression.[30]

A 2-step AI model was developed to improve the diagnosis of BE dysplasia (Figure 2). This consisted of an object detection model to scan whole slide images, identifying regions of interest, and making first-pass dysplasia grade predictions, which then were fed into a CNN model for secondary predictions.[31] The overall model combined both reads, with the highest congruent read as the final pathology read. After training on 368 whole slide images (all of which were graded and annotated by expert GI pathologists), a test set of 70 slides yielded sensitivity, specificity, and F1 score (measure of precision and recall) of greater than 80% each for the diagnosis of nondysplastic, low-grade, and high-grade dysplasia, a considerable improvement compared with the agreement as assessed in prior studies.[29,30] External validation of this model is

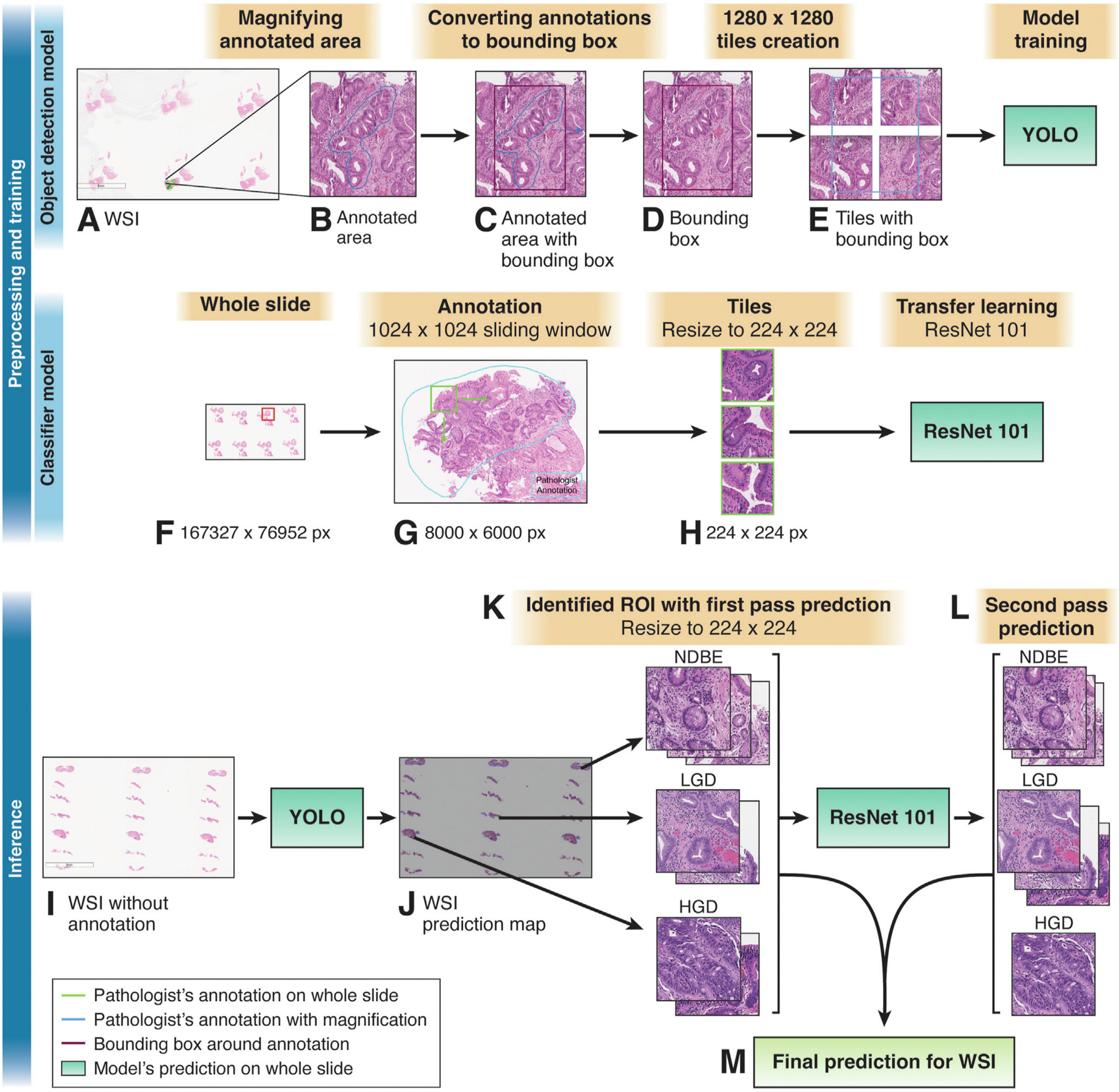
Preprocessing and training
Object detection model
Magnifying annotated area
Converting annotations to bounding box
1280 x 1280 tiles creation
Model training
YOLO
A WSI
B Annotated area
C Annotated area with bounding box
D Bounding box
E Tiles with bounding box
Classifier model
Whole slide
Annotation 1024 x 1024 sliding window
Tiles Resize to 224 x 224
Transfer learning ResNet 101
Pathologist Annotation
ResNet 101
F 167327 x 76952 px
G 8000 x 6000 px
H 224 x 224 px
Inference
K Identified ROI with first pass predction Resize to 224 x 224
L Second pass prediction
NDBE
LGD
HGD
YOLO
ResNet 101
I WSI without annotation
J WSI prediction map
M Final prediction for WSI
Pathologist's annotation on whole slide
Pathologist's annotation with magnification
Bounding box around annotation
🟩 Model's prediction on whole slide

Figure 2. Summary of preprocessing and training of object detection and classifier models, as well as process for inference by ensemble model.[31] (*A–E*) Object detection model development is shown. (*I–M*) Classifier model development is shown. (*F–H*) Ensemble inference is shown. (*A*) Whole slide images showing areas of annotations by expert GI pathologists. (*B*) Magnification view of annotation. (*C*) Creation of bounding box around annotation. (*D*) Segmentation of bounding box into 1280 × 1280 tile components. (*E*) Training for object detection model. (*F*) Whole slide with area of annotation. (*G*) Close up view of annotated areas. (*H*) Resizing of annotations to 224 × 224 tile segments followed by training of classifier model. (*I*) Removal of pathology annotations from test set slides with first-pass object detection model assessment. (*J*) Regions of interest as identified by object detection model. (*K*) Resizing of identified regions of interest to 224 × 224 pixels for feeding into object detection mode. (*L*) Second-pass prediction with classifier model. (*M*) Final dysplasia grade prediction by ensemble model. HGD, high-grade dysplasia; LGD, low-grade dysplasia; NDBE, non-dysplastic Barrett's esophagus; px, pixel; ResNet, residual neural network; WSI, whole slide image; YOLO, you only look once. Reprinted with permission from Faghani et al.[31]

underway, but this highlights the ability of AI to improve diagnostic capabilities in cases in which this traditionally has been a difficult task. Such a program could be used as an adjunct to guide pathologist diagnosis. In community practice, in which dysplastic BE slides may not be seen commonly,[30] this could result in fewer dysplastic overcalls, potentially leading to a decrease in unnecessary therapies, whereas in academic centers, this could improve interobserver agreement and increase pathologist confidence in the dysplasia diagnosis.

Liver nodules and steatohepatitis. Differentiating among various hepatocellular lesions is challenging, particularly high-grade dysplastic nodules from well-differentiated hepatocellular carcinoma.[32] To overcome this limitation, whole slide images of pathology slides from various hepatic adenomas, dysplastic nodules, and carcinomas, as well as focal nodular hyperplasia, were used to train 4 CNN models.[33] On a test set of 264 patients, all 4 models showed sensitivities and specificities exceeding 96%, with AUCs in excess of 0.99. Whole slide image classification maps correlated with markers that distinguished the diagnosis, mirroring what was annotated by study pathologists.

The assessment of steatotic liver disease and progression over time relies on histopathologic review of liver biopsy slides.[34] This is a time-consuming process with poor reproducibility, even among expert hepatopathologists,[35] and represents an area in which AI could make a meaningful impact. Slides were obtained from 3 randomized controlled trials that recruited patients with advanced fibrosis attributable to nonalcoholic steatohepatitis (NASH), and careful pathologist review and grading was completed before study inclusion.[36,37] A CNN model was trained on more than 1200 slides and subsequently tested on more than 3000 slides.[38] This model showed high concordance in NASH features when compared with the original pathologist grading (exceeding 0.6 for steatosis and fibrosis, although with poorer concordance for hepatocyte ballooning). Furthermore, segmentation of slides, which is not efficiently possible with human review in the clinical setting, showed considerable heterogeneity in histopathologic findings over time in individual patients, as well as among patients with similar clinical fibrosis scores. The authors of this study therefore believe this model can increase the accuracy of NASH histopathologic diagnosis and predict fibrosis progression while identifying patients who may benefit from therapeutic options for management of NASH.

Improve Clinical/Endoscopic Access While Limiting Costs

Rapid on-site evaluation in endoscopic ultrasound. Endoscopic ultrasound is an important facet of diagnostic endoscopy, enabling minimally invasive sampling of abnormal lymph nodes and masses. However, given the small-caliber needles used for sampling in these cases, suboptimal tissue acquisition is a well-established limitation.[39] Although larger-bore needles and increasing the number of passes may mitigate this issue, the concept of rapid on-site evaluation (ROSE) by a cytotechnologist has gained favor in the past decade, enabling confirmation of adequate sampling in real time during the procedure. However, this can be time consuming, and resources to have this service readily available limits broad applicability.

Utilizing advances with deep learning image analysis tasks, a CNN model was trained on 467 images obtained from 51 patients, with pathologists annotating cancer cells, and classifying whether images were inadequate, negative for malignancy, atypical, neoplastic (benign or malignant), suspicious for neoplasia, or definitively malignant. This model achieved sensitivity and specificity for cancer exceeding 79%, with accuracy greater than 83%, on a test set of 467 images on an internal test set, with similar results when applied to an external test set consisting of 693 images.[40] Similar adjunctive ROSE deep learning models assessing pancreatic masses have achieved similar results with AUCs greater than 0.90, surpassing that of trained endoscopists while nearing the performance of cytopathologists.[41]

These results support the feasibility of incorporating an AI cytology model into endoscopic ultrasound practice. This would enable widespread adoption of ROSE capabilities, particularly in resource-poor locations where an on-call cytopathologist may not be readily available. This also may improve clinical efficacy given the speed with which such samples can be analyzed.

Predict Prognosis

Barrett's esophagus and esophageal cancer. As previously described, BE is the only precursor lesion of EAC, which has poor morbidity and high mortality. However, although the risk of malignant progression in BE is increased significantly compared with the general population, the overall rate of progression is low, particularly in those with no dysplasia.[42] Progression prediction models, based on clinical factors alone, perform modestly,[43] resulting in a critical need to accurately identify patients who are truly at high risk of malignant progression and warrant more intense surveillance or consideration for ablation.

The TissueCypher assay (Castle Biosciences, Inc, Pittsburgh, PA) is an AI-powered digital platform for whole slide imaging using multichannel fluorescence, image object segmentation, as well as high-dimensional biomarker and morphology feature measurement, which is integrated with clinical data to prognosticate BE progression.[44,45] Samples are stratified into high-risk, intermediate-risk, and low-risk (for progression) classes based on progression risk over the next 5 years. In a pooled analysis of 552 patients, clinical variables predicting incident progression (defined as development of high-grade dysplasia/EAC 12 months or later from BE diagnosis), yielded a c-statistic of 0.68, whereas the addition of the TissueCypher assay risk assessment increased this significantly to 0.75, suggesting value to the use of this assay over the use of clinical variables alone.[46] Furthermore, nondysplastic patients in the high-risk TissueCypher category had an odds ratio of 14.3 for the risk of progression. This tool can assist clinicians with identifying high-risk patients who may benefit from more intense endoscopic surveillance or could consider ablation. TissueCypher is commercially available and the

Centers for Medicare and Medicaid Services has approved an Advanced Diagnostic Laboratory Test code for it.

AI may be capable of predicting prognosis in esophageal adenocarcinoma. A model was trained on diagnostic histology slides from 67 patients with gastroesophageal junction adenocarcinoma, and accurately distinguished patients who responded well to neoadjuvant therapy based on comparison of pretherapy with posttherapy positron emission tomography (at least 35% decrease in standardized uptake value maximum) with an accuracy of 0.78 ($P < .01$).[47]

Hepatocellular carcinoma. Hepatocellular carcinoma (HCC) is a lethal malignancy and ranks as one of the top causes of cancer-related mortality worldwide.[48] Prognostic factors on HCC tissue include microvascular invasion and tumor differentiation, although more subtle features such as tertiary lymphoid structures and vessels encapsulating tumor clusters also have been described.[49,50] These features may impact prognosis, but interpretation of these factors on histopathology is limited by subjectivity. However, AI may be able to formally segment these features and create a framework that may predict prognosis reliably based on biopsy specimens.

A CNN therefore was created to identify and map multiple elements that may be found on HCC histopathology, which are correlated with patient survival. This model was trained on 260 whole slide images, and using these data combined with survival, a tumor risk score was calculated to predict patient survival. This score was an independent predictor of survival in 2 validation sets, and exceeded prognostication capabilities when compared with clinical staging systems.[51] After segmentation, the features associated strongly with higher risk scores were sinusoidal capillarization, prominent nucleoli and nuclear envelope, and infiltrating inflammatory cells, providing a strong plausibility for the biologic basis of this model. Similar models also have shown good correlation of histopathologic findings with survival in HCC.[52]

Colorectal cancer. Colorectal cancer is a leading cause of cancer-related morbidity and mortality worldwide, and based on stage at diagnosis, therapeutic options include surgical resection, chemotherapy, and radiotherapy.

Quantification of histologic features may identify features that increase the risk of poor prognosis or could be combined with clinical factors to assess risk at the individual patient level. A quantitative model to segment features on colorectal cancer slides was trained on more than 550 annotated slides, with results combined with clinical and immunohistochemical staining to create a prognostic model for recurrence. This was validated both internally on 483 slides and externally on 938 slides. This model accurately distinguished high risk from low risk of disease recurrence at a per stage level (eg, at stage III, hazard ratio, 2.24; 95% CI, 1.33–3.87 for high- vs low-risk disease, respectively), and also predicted recurrence between these groups (at 36 months, 32.7% high-risk vs 13.4% low-risk). Removal of the quantification AI model from the overall prognostic model resulted in a significant performance decrease.[53]

The presence of lymph node involvement is associated with poorer cancer-specific survival outcomes and higher rates of recurrence.[54] Predicting lymph node involvement based on initial histologic sections can help streamline patient care and prognosticate outcomes at the individual patient level. A CNN was trained and validated on more than 1500 digitized slides obtained from 2 large cancer trial databases, and externally validated on a subset of 582 slides. This model achieved a modest AUC of 61.2% on this external test set, showing the feasibility of a model predicting lymph node metastasis based on cancer histopathologic slides. A limitation of this study was the heterogenous source of tissue, from both endoscopic and surgical resections, which could impact generalizability of these findings, particularly if this model's aim is to predict lymph node metastasis at initial diagnosis.

Predicting the risk of metastasis at the time of surgery for locally advanced colorectal cancer may help inform therapeutic decisions such as the need for adjuvant chemoradiotherapy. To assess whether AI can predict metastatic risk on surgical resection specimens, a model was trained on 102 patient specimens from resected locally advanced colon cancer, segmenting slides according to various features including ratios of smooth muscle, inflammation, stroma, and necrosis, among others.[55] The authors found that the model's assessment of inflammation and smooth muscle ratios correlated with metastatic probability.[55] Another model trained on early stage colorectal cancer patients with an external validation cohort incorporating data from more than 1100 patients showed excellent prognostic ability, with those assessed as having poor prognostication having a hazard ratio of 3.04 ($P < .0001$) for the outcome of cancer-specific mortality.[56] Other studies have shown models can assess for clinically actionable mutations and microsatellite instability accurately.[57,58]

How Will Artificial Intelligence Be Incorporated Into Clinical Practice?

A literature review yielded thousands of AI models evaluating their use in GI endoscopy, pathology, imaging, and medical records systems, the majority of which are built and trained on open-source code. Therefore, AI-powered solutions represent a unique domain from a regulatory standpoint, not clearly falling into previous categories such as drug development or medical devices, which traditionally have marked advances in medicine.

Most histopathologic ML models fall into the category of software as a medical device given that these models analyze medical images and information typically shared between health care providers, may provide recommendations to a health care provider, and the basis for those recommendations is not provided. To that end, the FDA has issued guidance regarding AI and ML software as a medical device and created a Digital Health Center of Excellence to innovate the regulatory approval process and efficiently (but safely) guide the implementation of AI models in clinical practice.[59]

Obtaining FDA approval for clinical use of an AI model is a rigorous process, but with careful design of algorithms and trials, approval can be granted using either the de novo pathway for novel AI models or via a 510(k) review for models that are substantially equivalent to an already existing one. The FDA also is working on innovative ways to expedite approval of medical devices, including vetting developers (and not individual devices), allowing these developers to forego premarket review for all devices they create, and also creating a framework for adaptive learning, wherein models can learn and adjust in real time based on clinical data it receives during regular use.[60]

Acknowledging the open-source nature of many of the models on which AI algorithms are based, it can be difficult to strike a balance between protection of intellectual property and commercialization of products. Innovators must work closely with their institutions or private funding partners when developing models to ensure that distribution of any profits of a product is determined early in the process.

Through July 2023, the FDA has approved more than 500 AI models for clinical use. In practice, will this result in the need to download dozens of models, or will a model package be delivered from which a singular program could be used to answer dozens of clinically relevant questions? What will ownership of the models look like in the future? How will such programs fit into the many different software/hardware programs and communicate across these devices as used in everyday practice? Furthermore, will models assessing similar but distinct questions (such as polyp identification based on endoscopy with tissue interpretation on whole slide images) learn from each other, as human beings do, to improve clinical knowledge? These questions are important, and the next few years hopefully will shed light on them.

From a practice perspective, the impact that AI will have on clinical jobs and compensation is unclear. Such algorithms could save time for clinicians to focus on other tasks. It is unlikely that AI will replace physicians, but how AI is used may impact physician compensation and potentially job prospects. Although we can only hypothesize what impact AI will have in the future, physicians must recognize that these models will have some impact on their day-to-day work in the near future and must be ready to adapt to their clinical practice, or risk being left in the dust.

What Are the Pitfalls of Artificial Intelligence–Powered Digital Histology?

Although we have examined several exciting areas in which the incorporation of AI may affect clinical practice favorably, as with any new technology, there are some limitations that may temper our enthusiasm for widespread adoption into practice. Given the rapid advancement of this technology, there are issues that are easily identifiable (although solutions may not be as forthcoming), but also concerns that we may not anticipate until AI is widely implemented in practice. For a summary of these challenges and potential solutions, please refer to Table 2.

Table 2. Challenges and Potential Solutions for Incorporating Artificial Intelligence Into Clinical and Research Practice

Challenge	Possible solutions
Data annotation	Human–computer interaction techniques (eg, active learning while annotating) Collaboration among subject matter experts to create publicly available data sets Using objective ground truths (labels)
Bias	Finding, learning, and applying techniques to be fair such as relying on multiple sources of data, avoiding unrepresentative data sets, overfitting hyperparameters
Heterogeneity of medical data	Preprocessing techniques, such as co-registration and color normalization
Variety of clinical tasks	Task-specific modeling to develop goal-directed models for each clinical task Developing an artificial general intelligence–based model for pathology detection rather than disease classification
Patient privacy	Generating synthetic data sets based on real data to protect patient information Using federated learning (enabling data to remain local without need for centralization of data, which can lead to security concerns)
Transparency	Using explanation techniques, such as saliency and attention maps
Reproducibility	Sharing code or providing detailed model descriptions
Uncertainty of model predictions	Incorporating uncertainty quantification techniques into model development

A Model Is Only as Good as What It Is Trained On

Many AI models have been developed to answer problems that arise when the histopathologic diagnosis is difficult to make, such as in the case of dysplastic BE[31] and grading of steatohepatitis.[38] Models typically use supervised learning, which requires accurate and reliable annotated imaging data, based on expert pathologist annotations; however, manual labeling is costly and time consuming, and it is well established that histopathologic interpretation can vary even among expert pathologists, leading to the possibility that these annotations may not be truly representative of the ground-truth.[29] This problem is magnified further when the diagnosis may rely on subjective criteria (eg, degree of inflammation as mild, moderate, or severe). Bias significantly affects data handling, model development, and slide evaluation.[61–63] As such, the ground-truth itself may not be based on a concrete foundation. Because a model may be only as good as the tissue that it is trained on, we may not fully appreciate a substantial improvement in diagnostic capabilities compared with our best human pathologists because the intangible difficulties in making these diagnoses are not accounted for by the model. This may explain the noted decrease in model performance seen on most external validation results. In addition, variation in stain characteristics (H&E or other special stains) also may influence AI model performance developed on slides from a single institution, although models to standardize stain characteristics have been developed.[64] The heterogeneity of medical data, variety of clinical tasks, patient privacy concerns, and algorithm trustworthiness further must be considered before widespread AI adoption.

To address these issues, several solutions come to mind. The ability to input long-term outcomes data when training models may enable such models to pick up on intangible factors that could better predict prognosis than currently used diagnostic models. For example, in BE histology, inputting into an AI model (clinical) factors that influence the risk of progression to EAC data could impact the way the model assigns a grade of dysplasia. Furthermore, models may need to report some form of explainability so clinicians can better understand model predictions. Models also may use segmentation of features (such as defining the exact degree of inflammation or a certain cell type, which may be too mundane and time consuming for human beings) to better understand the various degrees of pathologic processes. This would require significant upfront effort from an annotation standpoint but is incorporated commonly into model training today.

Concerns Regarding the Role of Artificial Intelligence in Digital Histology

We have discussed how AI can augment digital pathology, but what exactly does this mean? Many pathologists and clinicians would be uncomfortable without manual (human) overview of AI work, at least during the initial incorporation of AI into practice. However, removal of tedious, repetitive tasks is an area conducive for AI involvement. It is unclear if AI should be used as an adjunct, in which suspicious areas are highlighted for manual review (as in the case of WATS-3D),[65] should act as a filter enabling pathologists to focus on abnormalities (as in the case of colorectal biopsy specimens in which the program can sort out normal tissue not requiring further evaluation),[26] or should be used independently with occasional human review for quality assurance. Multiple GI and pathology societies have convened task forces to better address how exactly AI fits into clinical practice.[66,67]

At a more philosophical level, the introduction of AI highlights several moral and ethical concerns. With the expectation that AI will improve clinical outcomes, it may be expected that such tools are distributed equitably for all patients. However, the development of these algorithms requires significant computational resources, including space to house servers, connectivity frameworks, and a technically proficient workforce for upkeep of infrastructure. It remains unclear as to who will bear the brunt of these costs and whether the economics of AI price-out resource-poor nations and/or patients. Addressing these concerns as a clinical community is critical.

Conclusions

This is certainly an exciting time to be a practicing gastroenterologist or pathologist, and the incorporation of AI in digital histology is likely to dramatically affect patient care in a favorable manner. We have outlined several high-impact use cases for AI technology in GI pathology but acknowledge that limitations must be addressed before commercialization and during widespread adoption.

References

1. Busnatu Ș, Niculescu AG, Bolocan A, et al. Clinical applications of artificial intelligence-an updated overview. J Clin Med 2022;11:2265.
2. Jungmann SM, Klan T, Kuhn S, et al. Accuracy of a chatbot (Ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. JMIR Form Res 2019;3:e13863.
3. Semigran HL, Linder JA, Gidengil C, et al. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015;351:h3480.
4. Wasserlauf J, Vogel K, Whisler C, et al. Accuracy of the Apple watch for detection of AF: a multicenter experience. J Cardiovasc Electrophysiol 2023;34:1103–1107.
5. Ninh AQ. DocBot: a novel clinical decision support algorithm. Chicago, IL: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014:6290–6293.
6. SimilarityIndex™ | Hospitals. 2022. 2023. Accessed November 20, 2023. https://datalab.trillianthealth.com/similarity-index/hospitals/2022

7. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. FDA, 2023. Accessed November 20, 2023. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
8. FDA allows marketing of first whole slide imaging system for digital pathology. FDA, 2017. Accessed November 20, 2023. https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-whole-slide-imaging-system-digital-pathology
9. McCorduck P, Cfe C. Machines who think: a personal inquiry into the history and prospects of artificial intelligence. CRC Press, 2004.
10. López-Úbeda P, Martín-Noguerol T, Juluru K, et al. Natural language processing in radiology: update on clinical applications. J Am Coll Radiol 2022;19:1271–1285.
11. Robboy SJ, Gupta S, Crawford JM, et al. The pathologist workforce in the United States: II. An interactive modeling tool for analyzing future qualitative and quantitative staffing demands for services. Arch Pathol Lab Med 2015;139:1413–1430.
12. Shaheen NJ, Falk GW, Iyer PG, et al. Diagnosis and management of Barrett's esophagus: an updated ACG guideline. Am J Gastroenterol 2022;117:559–587.
13. Fitzgerald RC, di Pietro M, Ragunath K, et al. British Society of Gastroenterology guidelines on the diagnosis and management of Barrett's oesophagus. Gut 2014;63:7–42.
14. Reid BJ, Weinstein WM, Lewin KJ, et al. Endoscopic biopsy can detect high-grade dysplasia or early adenocarcinoma in Barrett's esophagus without grossly recognizable neoplastic lesions. Gastroenterology 1988;94:81–90.
15. Vennalaganti PR, Kaul V, Wang KK, et al. Increased detection of Barrett's esophagus-associated neoplasia using wide-area trans-epithelial sampling: a multicenter, prospective, randomized trial. Gastrointest Endosc 2018;87:348–355.
16. Codipilly DC, Krishna Chandar A, Wang KK, et al. Wide-area transepithelial sampling for dysplasia detection in Barrett's esophagus: a systematic review and meta-analysis. Gastrointest Endosc 2022;95:51–59.e7.
17. Collins MH, Martin LJ, Alexander ES, et al. Newly developed and validated eosinophilic esophagitis histology scoring system and evidence that it outperforms peak eosinophil count for disease diagnosis and monitoring. Dis Esophagus 2017;30:1–8.
18. Ricaurte Archila L, Smith L, Sihvo HK, et al. Performance of an artificial intelligence model for recognition and quantitation of histologic features of eosinophilic esophagitis on biopsy samples. Mod Pathol 2023;36:100285.
19. Klein S, Gildenblat J, Ihle MA, et al. Deep learning for sensitive detection of Helicobacter pylori in gastric biopsies. BMC Gastroenterol 2020;20:417.
20. Turner D, Ricciuto A, Lewis A, et al. STRIDE-II: an update on the Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE) initiative of the International Organization for the Study of IBD (IOIBD): determining therapeutic goals for treat-to-target strategies in IBD. Gastroenterology 2021;160:1570–1583.
21. Römkens TE, Kranenburg P, Tilburg Av, et al. Assessment of histological remission in ulcerative colitis: discrepancies between daily practice and expert opinion. J Crohns Colitis 2018;12:425–431.
22. Iacucci M, Parigi TL, Del Amor R, et al. Artificial intelligence enabled histological prediction of remission or activity and clinical outcomes in ulcerative colitis. Gastroenterology 2023;164:1180–1188.e2.
23. Siegel RL, Miller KD, Fuchs HE, et al. Cancer statistics, 2022. CA Cancer J Clin 2022;72:7–33.
24. Force UPST. Screening for colorectal cancer: US Preventive Services Task Force recommendation statement. JAMA 2021;325:1965–1977.
25. Joseph DA, Meester RG, Zauber AG, et al. Colorectal cancer screening: estimated future colonoscopy need and current volume and capacity. Cancer 2016;122:2479–2486.
26. Graham S, Minhas F, Bilal M, et al. Screening of normal endoscopic large bowel biopsies with interpretable graph learning: a retrospective study. Gut 2023;32:9512.
27. Wei JW, Suriawinata AA, Vaickus LJ, et al. Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. JAMA Netw Open 2020;3:e203398.
28. Nasir-Moin M, Suriawinata AA, Ren B, et al. Evaluation of an artificial intelligence–augmented digital system for histologic classification of colorectal polyps. JAMA Netw Open 2021;4:e2135271.
29. Vennalaganti P, Kanakadandi V, Goldblum JR, et al. Discordance among pathologists in the United States and Europe in diagnosis of low-grade dysplasia for patients with Barrett's esophagus. Gastroenterology 2017;152:564–570.e4.
30. Curvers WL, ten Kate FJ, Krishnadath KK, et al. Low-grade dysplasia in Barrett's esophagus: overdiagnosed and underestimated. Am J Gastroenterol 2010;105:1523–1530.
31. Faghani S, Codipilly DC, David V, et al. Development of a deep learning model for the histologic diagnosis of dysplasia in Barrett's esophagus. Gastrointest Endosc 2022;96:918–925.e3.
32. International Consensus Group for Hepatocellular Neoplasia. Pathologic diagnosis of early hepatocellular carcinoma: a report of the international consensus group for hepatocellular neoplasia. Hepatology 2009;49:658–664.
33. Cheng N, Ren Y, Zhou J, et al. Deep learning-based classification of hepatocellular nodular lesions on whole-slide histopathologic images. Gastroenterology 2022;162:1948–1961.e7.
34. Kleiner DE, Brunt EM, Van Natta M, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology 2005;41:1313–1321.
35. Merriman RB, Ferrell LD, Patti MG, et al. Correlation of paired liver biopsies in morbidly obese patients with suspected nonalcoholic fatty liver disease. Hepatology 2006;44:874–880.
36. Harrison SA, Wong VW-S, Okanoue T, et al. Selonsertib for patients with bridging fibrosis or compensated cirrhosis due to NASH: results from randomized phase III STELLAR trials. J Hepatol 2020;73:26–39.
37. Loomba R, Noureddin M, Kowdley KV, et al. Combination therapies including cilofexor and firsocostat for bridging fibrosis and cirrhosis attributable to NASH. Hepatology 2021;73:625–643.
38. Taylor-Weiner A, Pokkalla H, Han L, et al. A machine learning approach enables quantitative measurement of liver histology and disease monitoring in NASH. Hepatology 2021;74:133–147.
39. Schmidt RL, Witt BL, Matynia AP, et al. Rapid on-site evaluation increases endoscopic ultrasound-guided fine-needle aspiration adequacy for pancreatic lesions. Dig Dis Sci 2013;58:872–882.
40. Lin R, Sheng LP, Han CQ, et al. Application of artificial intelligence to digital-rapid on-site cytopathology evaluation during endoscopic ultrasound-guided fine needle aspiration: a proof-of-concept study. J Gastroenterol Hepatol 2022;38:883–887.

41. Zhang S, Zhou Y, Tang D, et al. A deep learning-based segmentation system for rapid onsite cytologic pathology evaluation of pancreatic masses: a retrospective, multicenter, diagnostic study. EBioMedicine 2022;80:104022.

42. Hvid-Jensen F, Pedersen L, Drewes AM, et al. Incidence of adenocarcinoma among patients with Barrett's esophagus. N Engl J Med 2011;365:1375–1383.

43. Rubenstein JH, McConnell D, Waljee AK, et al. Validation and comparison of tools for selecting individuals to screen for Barrett's esophagus and early neoplasia. Gastroenterology 2020; 158:2082–2092.

44. Prichard JW, Davison JM, Campbell BB, et al. TissueCypher(TM): a systems biology approach to anatomic pathology. J Pathol Inform 2015;6:48.

45. Gehrung M, Crispin-Ortuzar M, Berman AG, et al. Triage-driven diagnosis of Barrett's esophagus for early detection of esophageal adenocarcinoma using deep learning. Nat Med 2021; 27:833–841.

46. Iyer PG, Codipilly DC, Chandar AK, et al. Prediction of progression in Barrett's esophagus using a tissue systems pathology test: a pooled analysis of international multicenter studies. Clin Gastroenterol Hepatol 2022;20:2772–2779.e8.

47. Hörst F, Ting S, Liffers S-T, et al. Histology-based prediction of therapy response to neoadjuvant chemotherapy for esophageal and esophagogastric junction adenocarcinomas using deep learning. JCO Clin Cancer Inform 2023;7:e2300038.

48. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–249.

49. Calderaro J, Petitprez F, Becht E, et al. Intra-tumoral tertiary lymphoid structures are associated with a low risk of early recurrence of hepatocellular carcinoma. J Hepatol 2019;70:58–65.

50. Renne SL, Woo HY, Allegra S, et al. Vessels encapsulating tumor clusters (VETC) is a powerful predictor of aggressive hepatocellular carcinoma. Hepatology 2020;71:183–195.

51. Shi J-Y, Wang X, Ding G-Y, et al. Exploring prognostic indicators in the pathological images of hepatocellular carcinoma based on deep learning. Gut 2021;70:951–961.

52. Saillard C, Schmauch B, Laifa O, et al. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. Hepatology 2020;72:2000–2013.

53. Pai RK, Banerjee I, Shivji S, et al. Quantitative pathologic analysis of digitized images of colorectal carcinoma improves prediction of recurrence-free survival. Gastroenterology 2022; 163:1531–1546.e8.

54. Kim HJ, Choi GS. Clinical implications of lymph node metastasis in colorectal cancer: current status and future perspectives. Ann Coloproctol 2019;35:109.

55. Sirinukunwattana K, Snead D, Epstein D, et al. Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. Sci Rep 2018;8:13692.

56. Skrede O-J, De Raedt S, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet 2020;395:350–360.

57. Fu Y, Jung AW, Torne RV, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat Cancer 2020;1:800–810.

58. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer 2020;1:789–799.

59. Artificial intelligence and machine learning in software as a medical device. US FDA, 2021; 2023. Accessed November 20, 2023. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

60. How FDA regulates artificial intelligence in medical products. Pew, 2021; 2023. Accessed November 20, 2023. https://www.pewtrusts.org/-/media/assets/2021/08/ai_medicalproducts_issuebrief_final.pdf

61. Faghani S, Khosravi B, Zhang K, et al. Mitigating bias in radiology machine learning: 3. Performance metrics. Radiol Artif Intell 2022;4:e220061.

62. Rouzrokh P, Khosravi B, Faghani S, et al. Mitigating bias in radiology machine learning: 1. Data handling. Radiol Artif Intell 2022;4:e210290.

63. Zhang J, Li J. Mitigating bias and error in machine learning to protect sports data. Comput Intell Neurosci 2022;2022:4777010.

64. Michielli N, Caputo A, Scotto M, et al. Stain normalization in digital pathology: clinical multi-center evaluation of image quality. J Pathol Inform 2022;13:100145.

65. Vennalaganti PR, Naag Kanakadandi V, Gross SA, et al. Interobserver agreement among pathologists using wide-area transepithelial sampling with computer-assisted analysis in patients with Barrett's esophagus. Am J Gastroenterol 2015;110:1257–1260.

66. Parasa S, Repici A, Berzin T, et al. Framework and metrics for the clinical use and implementation of artificial intelligence algorithms into endoscopy practice: recommendations from the American Society for Gastrointestinal Endoscopy Artificial Intelligence Task Force. Gastrointest Endosc 2023;97:815–824.e1.

67. Eloy C, Bychkov A, Pantanowitz L, et al. DPA-ESDIP-JSDP Task Force for worldwide adoption of digital pathology. J Pathol Inform 2021;12:51.

Correspondence

Address correspondence to: Prasad G. Iyer, MD, MSc, Barrett's Esophagus Unit, Division of Gastroenterology and Hepatology, 200 1st Street SW, Mayo Clinic, Rochester, Minnesota. e-mail: iyer.prasad@mayo.edu.

Conflicts of interest

These authors disclose the following: Bradley J. Erickson is an officer at FlowSIGMA, Inc, and Yunu, Inc; and Prasad G. Iyer has received research funding from Exact Sciences, Pentax Medical, and Cernostics, is a consultant for Exact Sciences, Medtronic, Ambu Medical, Symple Surgical, and Cernostics. The remaining authors disclose no conflicts.