80897_Auto_Edited.docx

WORD COUNT	6218	TIME SUBMITTED	05-DEC-2022 12:06PM
		PAPER ID	93628379

Name of Journal: *Artificial Intelligence in Gastroenterology*
Manuscript NO: 80897
Manuscript Type: REVIEW

Artificial intelligence applications in predicting the behavior of gastrointestinal cancers in pathology

Aysen Yavuz, Anil Alpsoy, Elif Ocak Gedik, Mennan Yigitcan Celik, Cumhur Ibrahim Bassorgun, Betul Unal, Gulsum Ozlem Elpek

Abstract

Recent research has provided a wealth of data supporting the application of artificial intelligence (AI)-based applications in routine pathology practice. Indeed, it is clear that these methods can significantly support an accurate and rapid diagnosis by eliminating errors, increasing reliability, and improving workflow. In addition, the effectiveness of AI in the pathological evaluation of prognostic parameters associated with behavior, course, and treatment in many types of tumors has also been noted. Regarding gastrointestinal system (GIS) cancers, the contribution of AI methods to pathological diagnosis has been investigated in many studies. On the other hand, studies focusing on AI applications in evaluating parameters to determine tumor behavior are relatively few. For this purpose, the potential of AI models has been studied over a broad spectrum, from tumor subtyping to the identification of new digital biomarkers. The capacity of AI to infer genetic alterations of cancer tissues from digital slides has been demonstrated. Although current data suggest the merit of AI-based approaches in assessing tumor behavior in GIS cancers, a wide range of challenges still need to be solved, from laboratory infrastructure to improving the robustness of algorithms, before incorporating AI applications into real-life GIS pathology practice. This review aims to present data from AI applications in evaluating pathological parameters related to the behavior of GIS

cancer with an overview of the opportunities and challenges encountered in implementing AI in pathology.

INTRODUCTION

Gastrointestinal (GIS) cancers, including tumors of the esophagus, stomach, colon, and rectum, are an important health problem worldwide. Although the incidence of esophageal cancer (EC) is relatively low, gastric cancer (GC) and colorectal cancer (CRC) are among the most common types of cancer (fifth and third, respectively)[1]. They are also responsible for a substantial proportion of cancer mortality, with GC being the third and CRC the second most common cause of cancer-related death[2]. Although various predictive and prognostic parameters are currently available, the mortality rates for patients with GIS cancer are, unfortunately, still very high[2]. It has been shown that rectifying this situation may depend on paving the way for more personalized treatment strategies that lead to a better prognosis and/or fewer treatment side effects[3,4]. Therefore, the meticulous and complete evaluation of patients to determine the appropriate treatment is critical.

In this context, in addition to providing a definitive diagnosis, the role of an accurate evaluation of pathological parameters related to the behavior and proper treatment of GIS tumors cannot be ignored. However, pathology, a morphology-based specialty, is susceptible to subjectivity regarding intraobserver and interobserver variations, particularly in oncology. That is why, in recent years, the search for more objective criteria to eliminate bias, as well as to reduce the growing workload and to contribute time-saving, has allowed the improvement of image analysis-based digital pathology (DP), which has an important place in modern pathological applications[5,6].

In particular, significant advances in slide scanner technology, which can rapidly digitize all pathological slides at high resolution whole slide images (WSIs), has enabled not only the analysis of a wide range of morphological parameters but also the detection of biomarkers/genetic changes in many types of tumors[7-9]. The ability of computer-

based analysis to detect prognostic and predictive markers from these images, depending on the fact that they are composed of number matrices containing a large amount of information that is not accessible to the human eye, has led to the adoption of artificial intelligence (AI) for DP[10,11]. Accordingly, the number of studies on AI applications associated with the diagnosis, follow-up, and treatment of many tumors has increased significantly over time. Regarding GIS, data from previous studies evaluating pathological prognostic parameters with various AI models suggest that using these methods may be beneficial. Unfortunately, these encouraging results have not overcome the wide range of challenges to be solved, from laboratory infrastructure to improving the robustness of algorithms, before incorporating AI applications into real-life pathology practice.

This review presented the applications of AI in the evaluation of pathological parameters related to the behavior of GIS cancer, along with a brief overview of the opportunities and challenges encountered in its implementation in pathology.

GENERAL VIEW OF AI IN PATHOLOGY LABORATORIES

In parallel with technological developments, the evolution of whole slide imaging (WSI) has provided remote diagnosis, consultation, and education[12-14]. In the recent past, it was suggested that the use of WSI is comparable to, or even better than, conventional microscopic examination for decision-making in pathology[15-17]. On the other hand, WSIs are also crucial in applying AI methods in pathological practice. They not only provide quick access to the archive without loss of image quality, but they can also render gigabit images, which are very difficult to process, suitable for processing by "tessellation"[18]. This preprocessing is based on cutting a large image into nonoverlapping smaller patches called "tiles," making them amenable to computational analysis. It should be noted that although some pathological studies use selected images captured manually with a camera, WSI is currently recommended as a standard for AI applications, especially in tumors where heterogeneity is frequent, such as those of the GIS[19].

To achieve reliable results with WSIs, many steps, from preserving the structure of the tissue to the preparation of sections, must be carried out with care in the pathological laboratory. In particular, it is imperative to evaluate and check slides for artifacts (tears, floating contamination, thickness) that have the potential to adversely affect digitization and, thus, AI applications[20,21]. However, it should be noted that even with optimal protocols and slide scanner standardization, the importance of color normalization to ensure consistency in WSI databases should not be overlooked, as it can affect the robustness of deep learning (DL) models. Accordingly, histogram-matching color transfer and spectral matching methods can be applied[22-24]. However, as these methods depend on the expertise of pathologists and are impractical for manual adjustment, various algorithms have been proposed by researchers capable of performing this normalization. Although promising results have been obtained, there is a need for future studies on the performance of AI models using color normalization systems[25,26].

The gradual evolution of traditional pathology into DP has led to the development of powerful and user-friendly WSI analysis software tools with the ability to manage substantial WSIs and metadata from different hardware manufacturers, as well as interactive drawing annotation capabilities to facilitate decision-making and reporting. Moreover, a significant proportion of them is freely available[27-29]. In addition, the high costs of hardware required for high-performance computation in software development have become more affordable, leading to the implementation of DP in major medical centers[16,30-32]. Increasing the number of centers capable of using DP will allow for the generation of large and high-quality WSI databases, enabling the acquisition of large datasets and the design of algorithms for AI. However, the requirement of a significant investment is still an obstacle to overcome for the widespread application of these technologies[33]. In addition, the problem of proprietary datasets persists, limiting the repeatability of the proposed methodologies and hindering advancement in this field.

As mentioned above, the ability of AI to extract meaningful information from images that the naked human eye cannot discriminate makes it an attractive tool in the field of image processing and analysis in pathology. Therefore, contemporary AI models have

evolved from expert systems to different types, such as machine learning (ML) and DL (Table 1). In brief, ML is a subtype of AI that provides a computer system to automatically learn and develop from datasets on its own and solve problems without explicit programming[34-36]. DL is a subfield of ML that employs sophisticated algorithmic structures inspired by the neural network of the human brain (artificial neural network, ANN) in which statistical models are established from input training data[37-39]. Therefore, DL requires large, annotated datasets to develop its algorithms. At present, the annotation of datasets is a complex task in model development[9,40]. In practice, the time-consuming and challenging nature of annotation, especially in systems where heterogeneous lesions are common, such as GIS, may affect the accuracy of the model being trained[41]. Another limitation is that the dataset obtained by a study group does not show the same performance when compared to external validation sets from other institutions. Recently, studies have been conducted to overcome the hindering properties of annotation[42-44]. It has also been suggested that the adoption of DP for diagnosis could indirectly facilitate the generation of valuable datasets for future algorithm development by enabling pathologists to describe areas of interest during evaluation and reporting[45].

It has often been emphasized that the validation of AI-based technologies requires an evidence-based approach[42,46]. This should also be considered in a laboratory-based medical specialty such as pathology. On the other hand, analyzing the performance of AI techniques to that of pathologists is a significant challenge regarding interobserver and interobserver heterogeneity. Currently, the problems related to establishing "ground truth" in AI methods should not be overlooked[40,47]. It should be noted that this requires repeated testing of the effectiveness and consistency of AI applications in many different patient populations. The relative lack of a validation cohort in developing AI-powered DP applications is also related to the possible drawbacks of sharing histopathological slides. Despite interobserver heterogeneity and variability in pathological assessment also demonstrating the uncertainty of "ground truth" in this regard, multi center assessments involving multiple pathologists and datasets may be the best way to overcome this obstacle.

Before the integration of AI into the pathology workflow, the need to validate its benefits and address ethical recommendations increases the importance of AI-based tools being transparent and interpretable, resulting in an increasing demand for more explainable AI models. In this respect, there is a dilemma about the application of AI. Because most algorithms developed use DL, ensemble methods called "black box" models to tackle multidimensional problems are very complex. However, more straightforward methods that are not complex are not powerful enough to achieve the expected results[48]. For this reason, model interpretability, ethical concerns, and potential regulatory barriers should also be considered in newly developed AI tools to meet these expectations.

AI IN THE PATHOLOGICAL DETERMINATION OF PRENEOPLASTIC LESIONS IN GIS

Barrett's esophagus

The majority of AI studies in EC consist of imaging studies. In pathology, there have been recent studies on the diagnosis of Barrett's esophagus (BE) and the evaluation of dysplasia in these lesions to predict the risk of EC[49,50]. A proposed attention-based deep NN framework for detecting BE and adenocarcinoma (ADC) was found to be reliable with a mean accuracy of 0.83[49]. Unlike existing methods based on the region of interest, this model is based on tissue-level annotations, suggesting that it may provide a new approach for applying DL in pathology. On the other hand, the fact that the study was performed in a single center and on a relatively small data set necessitates the development of the proposed model with further studies. Since trefoil factor 3 expression is the key finding of BE, a DL model (VGG16) using immunohistochemically stained sections showed significant adaptability, with an area under the curve (AUC) of 0.88[50]. Although the proposed approach reduced the pathologist workload by 57%, the underlying ML model still needs further optimization.

Colorectal polyp classification

In CRC, unlike GC, the classification of polyps is an important task to determine the risk of CRC and the future surveillance needs of patients[51]. In routine examinations, high-risk polyps are evaluated based on their histopathological features with considerable interobserver variability among pathologists[52,53]. However, a precise diagnosis of high-risk polyps is required for efficient and early detection of cancer. In addition, the recommendation for endoscopic screening of these lesions for an early diagnosis of CRC, especially in elderly individuals, increases the workload of daily pathology practice[54].

Therefore, AI applications have been developed to classify high-risk colorectal polyps and/or adenomas with high-grade dysplasia. In studies on the classification of these lesions and the identification of CRC, datasets of three to six specific categories and five models were used[55-62] (Table 2). Although most studies showed good performance with generally high AUCs and accuracies, because of the following restrictions, the evidence level of each model needed to be improved. The number of patches and WSIs that make up the datasets are different. Accordingly, in some studies, the number of datasets may affect the reliability of the results. In various studies, the annotation process is not delineated in detail. In addition, the fact that each model has a different focus and characteristics makes their comparison across studies impossible. One of the most striking examples of these studies is Korbar *et al*[56], where a DL model (ResNet-152) trained with over 400 WSIs showed a high overall accuracy in subtyping polyps. In another study, Wei *et al*[61], who ensembled five layers of ResNet, could classify these lesions with WSIs from a single institution, even in external datasets with a performance comparable to that of histopathological evaluation. This data indicates that further manual annotations by various qualified GI pathologists may be required to decrease classification problems in future AI systems for colorectal polyp detection.

AI in the pathological determination of tumor behavior in GIS

In this section relevant data on GC and CRC will be discussed. Unfortunately, no AI studies have identified the parameters that are important in determining tumor behavior and survival in EC. Similarly, studies of EC concerning molecular characterization have

not been found. Therefore, in EC, a tumor with extremely high mortality, it is clear that additional pathology studies are necessary to reveal the effectiveness of AI applications in predicting tumor behavior.

TUMOR SUBTYPING

Gastric cancer

Although nearly all GC are ADC, the clinicopathological features and behaviors show considerable variation depending on the histopathological diversity of tumor cells[63,64]. In recent years, it has been reported that the survival of patients with GC at the same stage differs significantly among the different subtypes. Therefore, accurate histopathological classification is critical in determining their prognosis, monitoring, and treatment.

GC is often classified based on the ADC differentiation grade, including well-differentiated ADC and poorly differentiated ADC. The grading depends on the presence or absence of glandular structure formation. ADCs are divided into intestinal and diffuse subtypes based on the Lauren classification[65]. While the diffuse form comprises a poorly differentiated type and signet ring cell carcinoma (SRCC), the intestinal type exhibits glands with papillae, tubules, or solid regions. Diffuse-type carcinomas are commonly confused with other nonneoplastic diseases. Because they usually consist of solitary dispersed cells in a desmoplastic stroma and inflammation.

In most of the reported studies, the adenocarcinoma differentiation grade is judged through manual identification by pathologists. Although there have been many studies on AI applications in the pathological diagnosis of GC in the recent past, there are few studies regarding tumor subclassification (Table 3). Yasuda *et al*[66] investigated the features and classification of GC tissues by using supervised ML algorithms. The results showed that this method reliably identifies morphological changes in tumors with different grades. Interestingly, PD-L1 expression levels have been found to serve as a morphological classification in hematoxylin and eosin (HE)-stained slides and correlate with histological grades. Therefore, quantitative analyses of tissue morphology may

reveal molecular alterations in malignancies, and molecular analyses may aid in the pathological evaluation of cancer tissues. In another study, four different DL models were used to classify GC into diffuse ADC *vs* other ADC subtypes[67]. From biopsy WSIs, the trained model performed well at identifying both poorly differentiated ADC and SRCC cells. The authors pointed out that while higher magnification can reduce the false positive rate in classification, applying an RNN model with a more comprehensive dataset yields good results even at low magnifications. Hybrid models such as StoHisNet have also distinguished tubular, mucinous, and papillary subtypes of GC. This model showed a higher performance for multiclassification of pathological images of GC than other CNN-based models[68]. Although the model performed well in the four classifications of gastric pathological images, the study group does not include SRCC and other types. Also, the inability of the supervised network in the study to use unlabeled data and the lack of information on which combination maximizes the performance of the model performance warrant further studies. More recently, Su *et al*[69] demonstrated that DL models constructed using a pre-trained ResNet-18 model based on ImageNet27 achieved tumor differentiation recognition or poorly differentiated ADC and well-differentiated ADC classes, respectively. Although these results suggest that AI may be useful in GC classification, the scarcity of data and the differences in classification parameters used in these studies make it difficult to come to any solid conclusions.

Recently, GC has also been classified by the Tumor Cancer Genome Atlas (TCGA) into four molecular subtypes that are also included in the latest World Health Organization classification: Epstein–Barr-virus (EBV)-positive (9%), microsatellite unstable (MSI) (22%), genomically stable (GS) (19%) and chromosomally unstable (CIN) (50%)[70,71]. The clinical significance of this classification comes from the fact that various factors, such as the prognosis and treatment response, differ among these subtypes[72,73]. In particular, among all subclasses of GC, tumors with MSI and positive EBV are associated with a better response to immunotherapy[72]. Consequently, recognizing these subtypes is crucial for categorizing patients who benefit from these treatments. Nevertheless, such classification requires the application of costly techniques, such as

immunohistochemistry, and molecular testing, such as polymerase chain reaction, into pathological practice.

On the other hand, these two types have known characteristic histopathological findings. While EBV-positive GCs show prominent infiltration of lymphocytes into the neoplastic epithelium and the stroma, MSI subtype shows significant lymphocytic infiltration, intestinal-type histology, and expanding growth characteristics[63,74,75]. Therefore, these morphological features could be used to make predictions about the molecular subtype. In recent years, it has been suggested that molecular findings can be detected with AI *via* WSIs from HE-stained sections produced for pathological assessment[76-78]. Various models have been applied for molecular subtyping of GIS cancers. However, most of these studies have been conducted on CRCs (see below), whereas relatively few studies are available for GC (Table 3). For the detection of GC subtypes, Muti *et al*[79] demonstrated that DL could detect MSI and EBV positivity independently from each other in GC directly from HE-stained tissues in multi center pooled cohorts. They observed a high classification performance for the detection of MSI and EBV status. The relatively limited number of cases with positive findings and the fact that the ground truth methods for MSI were developed in CRC are presented as potential limitations of this study. On the other hand, their findings align with previous observations[69,80,81]. In addition, large-scale and multicenter validation broadens their work, which has considerable potential for integration into clinical procedures, suggesting that the application of DL could be a substitute for molecular techniques in the classification of GC. Furthermore, because these two subtypes share common morphological features and they are immunotherapy-sensitive tumors, Hinata *et al*[82] combined MSI and EBV in DL models and found they had a higher detection accuracy. This finding has been interpreted based on the possibility that these subtypes have similar distinctive pathological features, such as abundant stromal lymphocytic infiltration and intraepithelial lymphocytosis. On the other hand, the use of tissue microarray and manual labeling of tumor regions for TCGA presented as sources of bias compared to whole tissue slides, given the heterogeneity of tumor tissue. It was also

emphasized that manual annotation by a pathologist might be a challenge to overcome by some weakly supervised methods (for example, attention-based deep multi instance learning) in the field of DL for the broad application of the proposed model.

Recently, a DL model called EBVNet that assists pathologists in predicting EBV from HE-stained slides has been introduced in GC[83]. The results suggested that human-machine dramatically enhances the diagnostic ability of both EBVNet and the pathologist. However, this study has some limitations regarding its retrospective evaluation of training and validation. Additionally, the logistic regression model applied in the assessment is still an indirect way to interpret the model. More importantly, as in many DL models, the EBVNet decision-making procedure by the neural network is nontransparent (black boxes). Since various methods have been proposed to solve black boxes in DL in the recent past, additional studies applying these methods will contribute to the determination of the molecular subtypes of AI models of GC[84-86]. In a more recent study, Flinner *et al*[87], in their study emphasizing the error-proneness of the morphological and staining methods used to determine GC subtypes for subclassification, found that DL could be more effective in this regard. On the other hand, they also pointed out that image tiles labeled with false ground truth associated with GC heterogeneity may reduce the accuracy of DL but this can be overcome by first experimentally defining the test data.

Recently, the feasibility of a DL approach has also been evaluated in the classification of GC for mutations in the CDH1, ERBB2, KRAS, PIK3CA, and TP53 genes[88]. High AUCs observed in both frozen and formalin-fixed tissues highlight that DL-based classifiers could predict the mutational status of these tumors. Although these results are promising for the application of AI to subtyping GC, additional studies are necessary, with further refinement of these methods.

Colorectal cancer

Similar to GC, molecular subtyping of CRC is essential for targeted treatment against critical oncogenic signaling pathways. CRCs are divided by molecular consensus into

four types (CMS): 1. CMS1: Tumors with MSI that have a good prognosis in non metastatic stages; CMS2: Tumors with intermediate prognosis exhibiting epithelial gene expression, activated WNT and MYC signaling; CMS3: Tumors with intermediate prognosis demonstrating metabolic dysregulations; CMS4: Tumors with a poor prognosis that possess transforming growth factor beta (TGF-b) activation[89-91]. The identification of CRC with MSI is paramount because this group is susceptible to immunomodulating therapies[92,93]. Although some findings, such as tissue architecture, growth pattern, cellular morphology, and distributions of tumor stroma ratio (TSR) and tumor microenvironment (TME) provide some clues about the subclassification of these tumors, molecular stratification of patients necessitates RNA analyses that are expensive and difficult to standardize[94-96]. Accordingly, some studies have investigated the contribution of AI to tumor subclassification from HE-stained tissue sections by DL models (Table 2). Sirinukunwattana *et al*[97] demonstrated that a CNN-based model could detect CMS subtypes. At the same time, they criticized the potential over fitting of the computational model to the training cohort as a limitation of the study. In a more recent study, Echle *et al*[98] developed a DL model in a large series of 8836 cases of CRC to predict MSI tumors. In the international validation of the study group, the algorithm achieved a high performance [area under the receiver operating curve (AUROC) of 0.96][80]. Other investigators have also reported similar results, pointing out the potential use of DL models for detecting molecular subtypes of CRC[77,99-101]. In a retrospective study, a DL pipeline method was developed based on experimental setups similar to previous studies[102]. Three models were used to predict mutation density (low *vs* high), MSI, CIN, and GpG island methylator phenotype. The mutated and wild-type BRAF, TP53, and KRAS types were also investigated. This method showed higher AUROCs for the prediction of hypermutation, MSI, CIN, BRAF, and TP53 compared to previously reported data, suggesting that AI methods may provide the stratification of patients with CRC for targeted therapies. However, further large-scale validations with multicenter datasets are required before their implementation in pathological practice.

LYMPH NODE METASTASIS

Gastric cancer

Another important parameter that predicts GC behavior and treatment is lymph node metastasis (LNM)[103]. However, identifying LNM is still a challenging and tedious task in pathological practice, making the implementation of AI an attractive tool to reduce the workload[104,105]. Although numerous studies have demonstrated that DL-based algorithms can detect metastatic lymph nodes in GC with a similar level of accuracy to human specialists, these algorithms have not yet been implemented into pathology practice[106-108]. (Table 3). The failure to integrate these algorithms is related to the characteristics of WSIs, the excessive effort required to apply the annotation, and the limited associated data. Recently, Huang *et al*[109] developed a weakly supervised end-to-end technique termed enhanced streaming CNN (ESCNN). Their results revealed that the routine pathological evaluation benefitted from the AI-assisted LN assessment workflow regarding review time, sensitivity, and consistency. On the other hand, AI-attributable false alarms that misled the pathologists on negative results led to a decrease in specificity from 94% to 84%, which needs more large-scale or multicenter studies to check the effectiveness of the workflow.

Colorectal cancer

Recent evidence indicates that features extracted by DL models from routine histologic slides can predict LNM in CRC[110-112] (Table 2). For example, Kwak *et al*[110] detected LNM by generating a score based on the ratio of peritumoral stroma to tumor tissue on a test set. In another study, the presence of LNM was detected with a model which segmented WSIs into areas such as tumor budding or poorly differentiated clusters[111]. More recently, Kiehl *et al*[112] performed an approach that uses DL-based image analysis (slide-based artificial intelligence predictor) in association with patient data to estimate LNM in CRC patients. Their results indicated that LNM could be predicted in patients with CRC through AI applications from histological slides to a similar level to using a classifier containing clinical data.

THE TUMOR STROMA RATIO, TUMOR MICROENVIRONMENT AND TUMOR BUDDING

Gastric cancer

In recent years, it has been shown that the TSR in many organ tumors is an important clue to the course of the disease. In particular, stromal dominance has been observed to be an independent prognostic factor in many tumors, including GIS[113,114]. However, TSRs are not included in pathology report protocols because of the lack of a standard procedure among different methodologies and a low reproducibility related to the high interobserver variation[115]. Recently, a DL pipeline has been introduced to facilitate the automated assessment of TSR in GC[116]. Although this model has been shown to be effective in detecting survival according to the low and high TSR rates in advanced GC, it was emphasized that some limitations, such as the nonautomatic selection of hot spots and the use of a single test, should be eliminated. Therefore, there is a need for many studies on the use of AI applications in TSR determination of GC.

In a recent study, a DL model determined the tumor-to-metastatic lymph node-area ratio in metastatic lymph nodes in patients with GC[106]. Statistical analysis also revealed that this ratio is an independent prognostic factor warranting further investigation.

Colorectal cancer

In CRC, recent studies have demonstrated that lymphocytes and fibroblasts profoundly shape the TME and significantly impact tumor behavior[117-119]. In addition, it has been shown that CRC may have a poor prognosis due to tumor budding (1-5 cells in the invasive area)[120]. In the literature, seven studies of AI methods have been identified to determine these parameters in a more objective and time-saving manner (Table 2). However, many of them used different methods. Three models focused on the classification of the cell types, such as epithelial, inflammatory, fibroblast, lymphocytes, and others (mucus, smooth muscle, normal mucosa, stroma, and cancer epithelium)[121-123]. In an elegant study, a DL algorithm was proposed for estimating the risk of distant

metastasis by analyzing the TME[123]. Cell detection and cell classification were evaluated in two CNNs used to build a cell network. In each tumor, a tissue phenotype signature was obtained by proportioning the area of tissue phenotypes to the total tissue area. Statistical analysis revealed that the connection frequency (CF) of the smooth muscle ratio, the CF of the inflammation ratio, and the appearance (AP) based on inflammation could independently estimate the development of distant metastasis. Distant metastasis-free survival analysis indicated that CF smooth muscle and AP inflammation ratios were potential prognosticators. Although the hazard ratios for CF of the smooth muscle ratio and AP inflammation were 2.11 and 0.39, respectively, the AUC values for distant metastasis prediction were 0.59 for the CF of the smooth muscle ratio and 0.64 for AP based on inflammation. As emphasized by the authors, specific immunohistochemical staining can improve the prediction of distant metastases by increasing the informative value of histological slides. Another limitation of this study is the small number of metastatic cases. Another recent study was performed to detect CD3- and CD8-positive immune cells on WSIs of slides stained by immunohistochemistry in a multicenter cohort by four different methods[124]. U-Net obtained the highest performance and highest agreement with manual evaluation (0.72), which was higher than that of pathologists (K=0.64), supporting that DL models are helpful for automatically detecting lymphocytes in immunohistochemically stained tissue sections.

In CRC, the automatic tumor budding evaluation on immunohistochemical pankeratin-stained slides revealed that the absolute number of buds per image was significantly correlated with manually segmented ground truth (R: 0,86)[120]. Interestingly, the number of spatial clusters of buds in hot spots was significantly correlated with the prognosis. In three studies, the impact of detecting the TSR or deep stroma score in CRC by DL algorithms was found to be an independent parameter to predict tumor behavior[115,121,125] (Table 2).

Recently, Zhao *et al*[126] demonstrated that the ratio of the mucinous component in the tumor area (MTR) quantified by AI is an independent prognostic factor in CRC. On the other hand, the most invasive part of primary tumors was selected for evaluation. As

noted by the authors, measuring the exact proportion and prognostic value of mucus in the entire tumor is still worthy of further investigation.

SURVIVAL OUTCOMES

Gastric cancer

Another continuing research topic is evaluating survival outcomes in GC with AI models[127-129] (Table 3). Recently, support vector machine (SVM), one of the popular algorithms in ML, has been applied to predict the survival of GC. Jiang *et al*[129] demonstrated that SVM could be useful in predicting the outcome and identifying patients with GC who might benefit from adjuvant therapy. In this study, the classifier incorporated patient gender, carcinoembryonic antigen levels, LNM, and the protein expression level of eight features, composed of CD3 invasive margin (IM), CD3 center of the tumor (CT), CD8IM, CD45ROCT, CD57IM, CD66bIM, CD68CT, and CD34. There were significant variations between the high- and low-GC-SVM classifiers. Recently, Huang *et al*[128] designed MIL-GC (a DL-based model) to predict overall survival (OS) in patients with GC. They observed C-indices of 0.728 and 0.671 in the training and internal validation sets, respectively. The external validation likewise exhibited strong prognostic prediction performance (C-index = 0.657), confirming the resilience of the two models. Furthermore, univariate and multivariate Cox analyses demonstrated that the risk score derived by MIL-GC has independent prognostic significance, indicating the potential of AI approaches to predict GC behavior. Additionally, tumor progression includes complex interactions between malignant cells and their surrounding microenvironment (TME)[130]. TME targeting and reprogramming is, in fact, can be a potential strategy to achieve antitumor effects in many cancers. Several AI studies involving the TME have recently demonstrated that these methods can determine the prognosis of GIS cancers. Regarding GC, Wang *et al*[131], suggested a graph NN-based solution, CellGraph Signature powered AI, for the digital staging of TME and the exact prediction of patient survival by combining and converting multiplexed immunohistochemistry (mIHC) images as Cell-Graphs. The survival prediction achieved outstanding model performance

for both binary and ternary classifications. Furthermore, survival analysis revealed that this method outperforms the AJCC 8th edition Tumor Node Metastasis staging system in discriminating both binary and ternary classes with statistical significance (*P* value < 0.0001), implying the effectiveness and advantages of such an AI-powered digital staging system in DP and precision oncology.

These data demonstrate that AI-based models allow prognosis prediction in GC. However, developing efficient models requires training on large sets reflecting scanning and staining protocols variability.

Colorectal cancer

Regarding prognostic evaluations from HE-stained slides by AI in CRC, some DL models have been developed for prognostication (Table 2). Bychkov *et al*[132] combined a CNN and a recurrent NN model to estimate the disease-specific five-year survival from tumor tissue microarray samples without tissue classification. The model classified patients into a low- or high-risk group (AUC of 0.69). This result was more significant than the AUC of the visual evaluation of the pathologist (AUC of 0.58) or the histological grade determined at the time of the original diagnosis (AUC of 0.57). However, an external dataset was not included. In another study by Skrede *et al*[133], diverse data from four different cohorts were used to develop an automatic prognostic marker to predict the outcome. The model included a CNN used to separate tumor tissue and two other CNN ensembles that identified individuals as having a favorable or poor survival. Patients were assigned as uncertain when the two CNN ensembles predicted different outcomes. In an external test group, the classifier was a strong predictor of survival. In addition, the output of the two CNN ensembles produced a strong predictive score related to patient outcome (AUC of 0.71). A generalization of this approach has been recommended, as an external test cohort from more than one medical center demonstrated similar hazard ratios.

Jiang *et al*[134], to achieve a shorter computational time, developed a hybrid model by synergizing ML algorithms with DL (InceptionResNetV2 and gradient boosting decision

machine classifier) to predict the survival of patients with stage III CRC. While the internal test sets constituted a Chinese cohort, external testing was performed on the TCGA cohort. They revealed that the model stratifies patients with stage III colon cancer into high- and low-risk recurrence and poor and favorable prognostic groups directly from tissue sections. These data suggest that the analysis of H-E-stained tissue samples by AI methods could serve as a digital prognostic biomarker in CRC. However, additional studies are warranted to support the evaluation of the performance of these methods in larger patient series.

OVERALL LIMITATIONS OF AI-BASED APPLICATIONS IN REAL-LIFE PRACTICE

In the literature, there are some frequently discussed topics considering the general challenges of AI such as identification of the clinical need, ethical considerations, funding, optimization of data-sets, annotation of the dataset, regulation, validation, and implementation[46].

Recognizing the actual clinical need and defining a potential solution is the first stage in developing the AI application. However, there can be an imbalance between the benefits in daily pathological practice and the total cost of its implementation. As a result, the market for a particular AI tool may be too tiny and it may not be profitable.

Although patients can provide permission for data to be used for studies, constructing AI models may have issues if commercial use is not approved[135]. In order to develop a framework for global data sharing, patient consent should include the possibility of its commercial use for product development[40].

Training on huge datasets is necessary for developing AI systems with high performance in digital pathology. Changes related to differences in fixation, tissue thickness, and variations in staining and scanning protocols encountered in preanalytical and analytical phases may influence data accuracy[136,137]. For example, it is difficult to convert a glass slide to WSI, and changing the hue of the slide could affect AI accuracy. Many AI algorithms have emerged for this purpose recently, including staining and color

features[138,139]. In addition, a number of algorithms are presented to optimize WSI quality. These algorithms identify areas of the highest quality and exclude areas that are out of focus or affected by artifacts[140,141].

Concerning the implementation of AI, to enable users to shift the daily routine practice in the pathology laboratory, from glass slides to WSIs, the first step is to install an institutional IT infrastructure. In addition to these changes in infrastructure, pathology residency training might need to be adjusted in accordance with the availability of this new tool. Preventing residents from relying completely on AI while also allowing them to benefit from it as a helping instrument would require fine balancing and planning prior to its installation[142].

Similar to other clinical tests, quality assurance is crucial, hence it is urgently necessary to develop a plan for external quality assurance for applications. Furthermore, laboratory workers should also be familiar with the quality management system.

Although some algorithms and automated AI models are thought to perform better than pathologists, pathologists will always be required to audit technology and control mechanisms in AI implementation[143].

CONCLUSION

In this review, we outlined the potential of AI applications for evaluating pathological parameters related to the behavior of GIS cancers. Current data suggest the merit of AI-based approaches in assessing tumor grading, subtyping, detection of metastasis, and prognosis in GC and CRC. In addition, these methods encourage biomarker discovery by revealing predictions that are impossible when using traditional visual methods. Regarding EC, there is still much room for improvement in developing AI models to predict the behavior of these tumors in pathology. On the other hand, the enormous potential of AI in improving workflows, eliminating simple errors, and increasing objectivity during pathological evaluations to determine the behavior of GIS cancers should motivate researchers to overcome the many remaining hurdles. In algorithm development, variations in imaging data, interobserver variability during

interpretations, model transparency, and interpretability are significant challenges to be solved. A large number of studies with external validation and quality controls implemented on large datasets are essential in meeting the standards of these methods. Thereby, AI applications that are practical, interpretable, manageable, and cost-effective can play a crucial role in the development of pathological evaluations to be performed in the prognosis and treatment of GIS tumors.

80897_Auto_Edited.docx

ORIGINALITY REPORT

3%
SIMILARITY INDEX

PRIMARY SOURCES

1	Sara Kuntz, Eva Krieghoff-Henning, Jakob N. Kather, Tanja Jutzi et al. "Gastrointestinal cancer classification and prognostication from histology using deep learning: Systematic review", European Journal of Cancer, 2021 Crossref	35 words — 1%
2	Chandavalli Ramappa Raghushaker, Jackson Rodrigues, Subramanya G Nayak, Satadru Ray et al. "Fluorescence and Photoacoustic Spectroscopy-Based Assessment of Mitochondrial Dysfunction in Oral Cancer Together with Machine Learning: A Pilot Study", Analytical Chemistry, 2021 Crossref	32 words — < 1%
3	www.wjgnet.com Internet	26 words — < 1%
4	www.nature.com Internet	18 words — < 1%
5	www.omicsdi.org Internet	17 words — < 1%
6	Maxime Chénard-Poirier, Elizabeth C. Smyth. "Immune Checkpoint Inhibitors in the Treatment of Gastroesophageal Cancer", Drugs, 2019 Crossref	16 words — < 1%

7	www.researchgate.net Internet	15 words — < 1%
8	link.springer.com Internet	12 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS