

# Endoscopy

## An Artificial Intelligence System for Distinguishing Between Gastrointestinal Stromal Tumors and Leiomyomas Using Endoscopic Ultrasonography (with video)

Xintian Yang, Han Wang, Qian Dong, Yonghong Xu, Hua Liu, Xiaoying Ma, Jing Yan, Qian Li, Chenyu Yang, Xiaoyu Li.

Affiliations below.

DOI: 10.1055/a-1476-8931

**Please cite this article as:** Yang X, Wang H, Dong Q et al. An Artificial Intelligence System for Distinguishing Between Gastrointestinal Stromal Tumors and Leiomyomas Using Endoscopic Ultrasonography (with video). Endoscopy 2021. doi: 10.1055/a-1476-8931

**Conflict of Interest:** The authors declare that they have no conflict of interest.

**This study was supported by** National Natural Science Foundation of China (<http://dx.doi.org/10.13039/501100001809>), Grant81802777, "Clinical medicine + X" scientific research project of Affiliated Hospital of Qingdao University

**Trial registration:** ChiCTR2000039322, Chinese Clinical Trial Registry (<http://www.chictr.org/>), Prospective diagnostic test

### Abstract:

#### Background

Gastrointestinal stromal tumors (GISTs) and gastrointestinal leiomyomas (GILs) are the most common subepithelial lesions (SELs). All GISTs have malignant potential; however, GILs are considered benign. Current imaging cannot effectively distinguish GISTs from GILs. We aimed to develop an artificial intelligence (AI) system to differentiate these tumors using endoscopic ultrasonography (EUS).

#### Methods

The AI system was based on EUS images of patients with histologically-confirmed GISTs or GILs. Participants from four centers were collected to develop and retrospectively evaluate an AI-based system. The system was used when endosonographers considered SELs as GISTs or GILs. We used the system in a multicenter prospective diagnostic test to clinically explore whether joint diagnoses by endosonographers and the AI system can distinguish between GISTs and GILs to improve the total diagnostic accuracy of SELs.

#### Results

The AI system was developed using 10439 EUS images from 752 participants with GISTs or GILs. In the prospective test, 132 participants were histologically-diagnosed (36 GISTs, 44 GILs, and 52 other types of SELs) among 508 consecutive subjects. Through joint diagnoses, the total accuracy of endosonographers in diagnosing the 132 histologically-confirmed participants increased from 69.7% (95% confidence interval [CI]: 61.4–76.9%) to 78.8% (95% CI: 71–84.9%;  $p=0.012$ ). The accuracy of endosonographers in diagnosing the 80 participants with GISTs or GILs increased from 73.8% (95% CI: 63.1–82.2%) to 88.8% (95% CI: 79.8–94.2%;  $p=0.012$ ).

#### Conclusions

We developed an AI-based EUS diagnostic system that can effectively distinguish GISTs from GILs and improve the diagnostic accuracy of SELs.

### Corresponding Author:

Han Wang, The Affiliated Hospital of Qingdao University, Department of Pathology, Shenzhen, China, wanghan\_125@163.com

**Affiliations:**

Xintian Yang, The Affiliated Hospital of Qingdao University, Department of Pediatric Surgery, Qingdao, China

Han Wang, The Affiliated Hospital of Qingdao University, Department of Pathology, Shenzhen, China

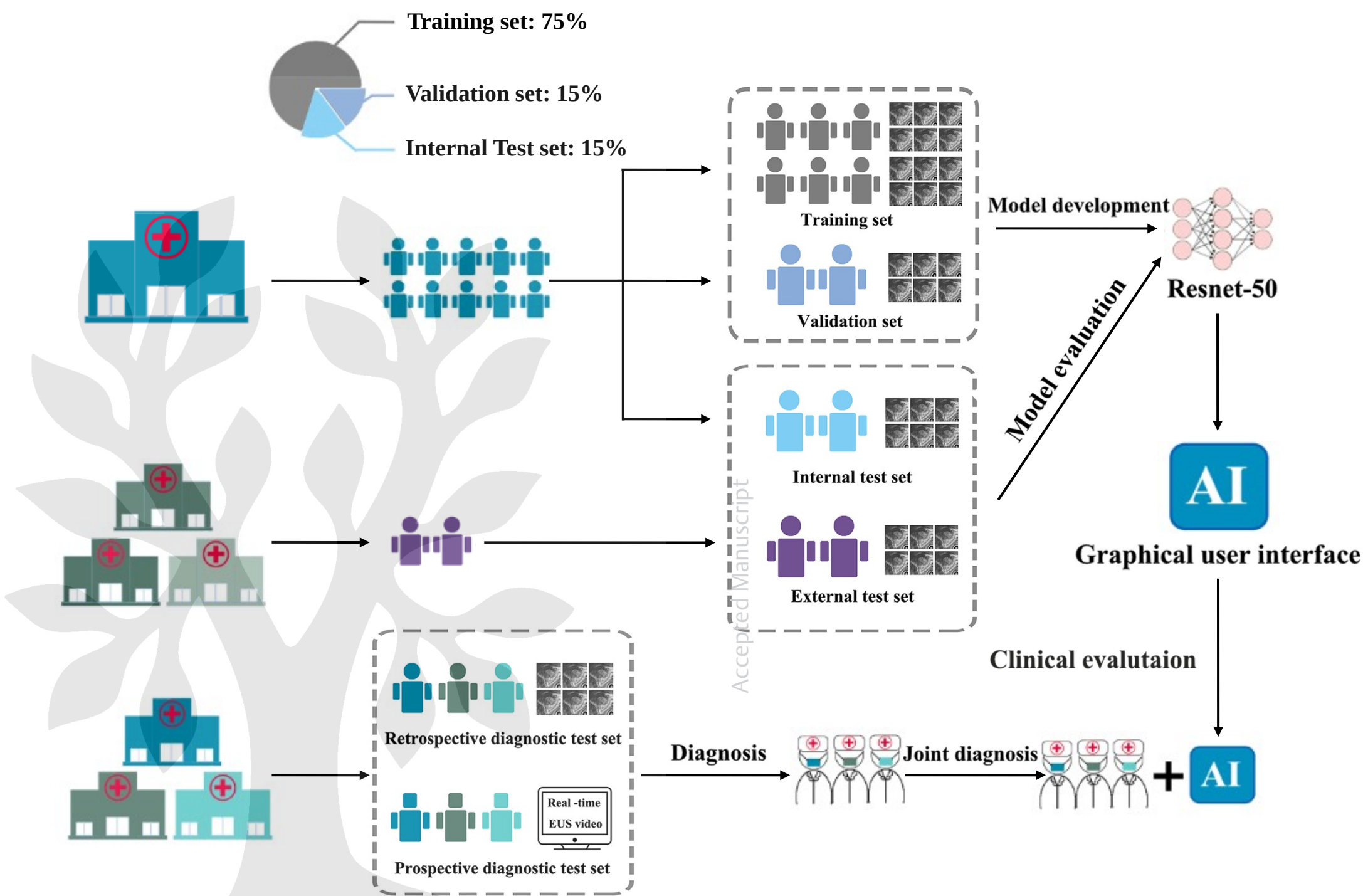
Qian Dong, The Affiliated Hospital of Qingdao University, Department of Pediatric Surgery, Qingdao, China

[...]

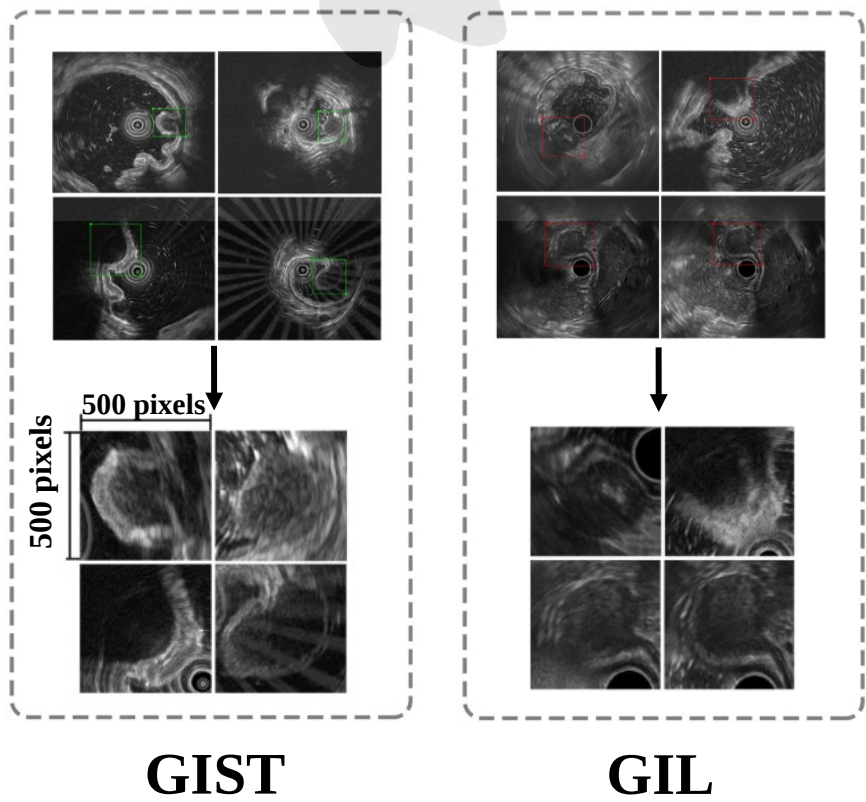
Xiaoyu Li, The Affiliated Hospital of Qingdao University, Department of Gastroenterology, Qingdao, China



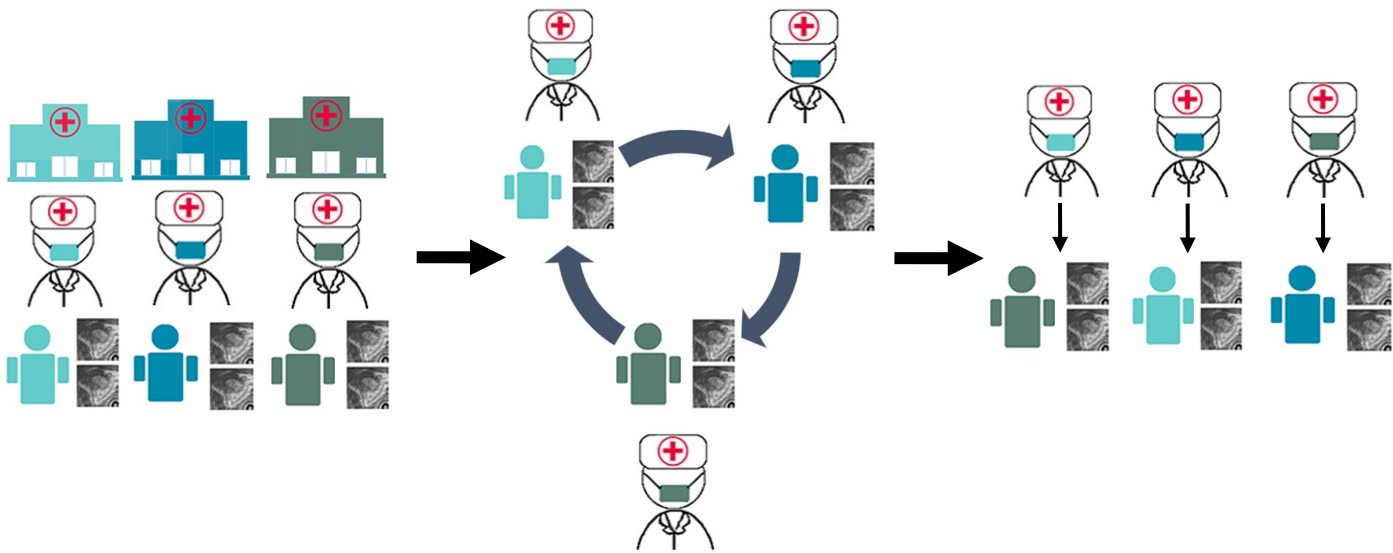
**A**

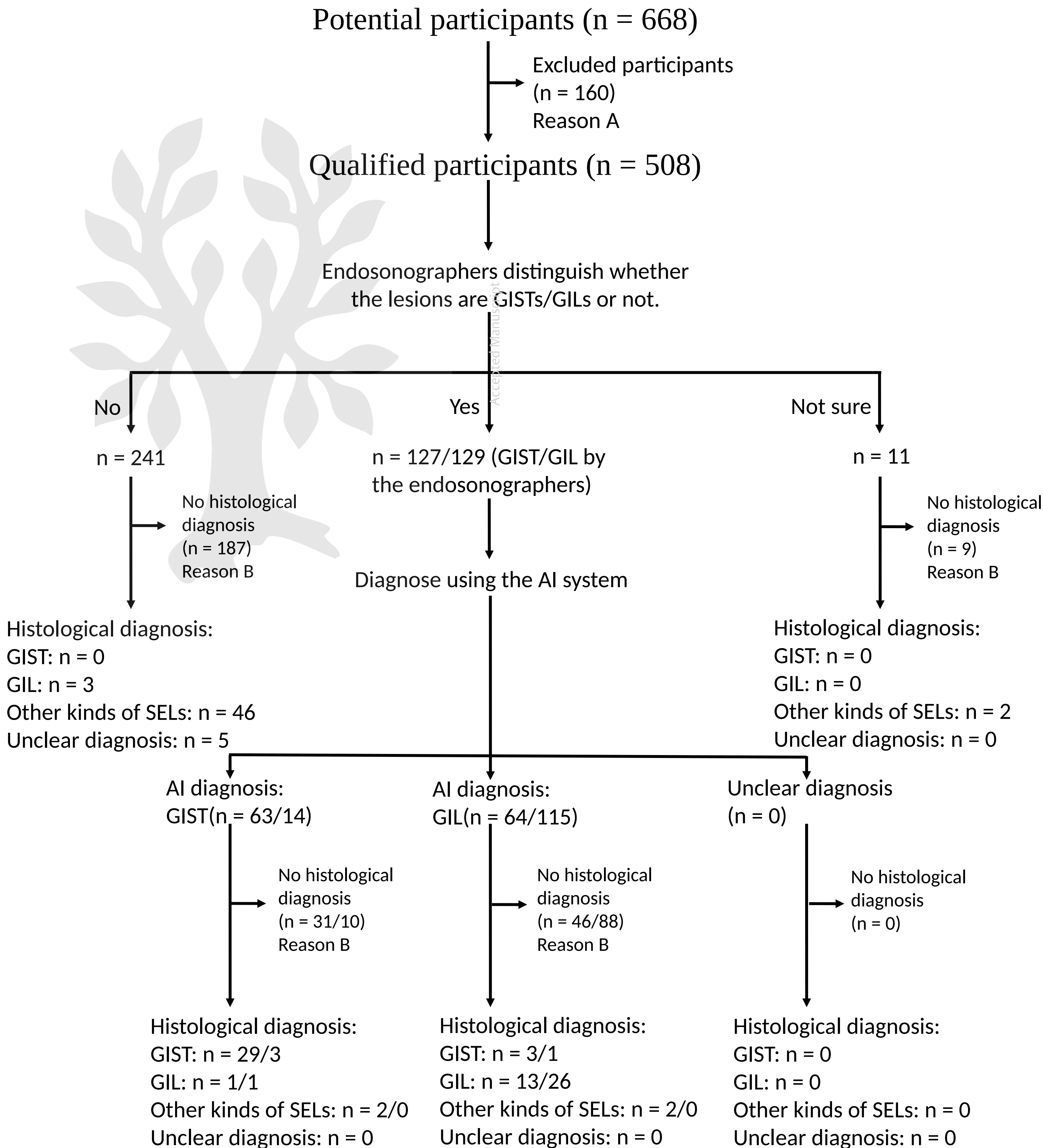


**B**



**C**





# An Artificial Intelligence System for Distinguishing Between Gastrointestinal Stromal Tumors and Leiomyomas Using Endoscopic Ultrasonography (with video)

## Authors and affiliations

Xintian Yang<sup>1,2</sup>, Han Wang<sup>3</sup>, Qian Dong<sup>1,2</sup>, Yonghong Xu<sup>4</sup>, Hua Liu<sup>4</sup>, Xiaoying Ma<sup>5</sup>, Jing Yan<sup>4</sup>, Qian Li<sup>4</sup>, Chenyu Yang<sup>1,2</sup>, Xiaoyu Li<sup>4\*</sup>

1. Department of Pediatric Surgery, The Affiliated Hospital of Qingdao University, Qingdao, China.

2. Shandong Key Laboratory of Digital Medicine and Computer Assisted Surgery, The Affiliated Hospital of Qingdao University, Qingdao, China.

3. Department of Pathology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, China

4. Department of Gastroenterology, The Affiliated Hospital of Qingdao University, Qingdao, China.

5. Department of Gastroenterology, Qingdao Municipal Hospital, Qingdao, China.

## Correspondence

Xiaoyu Li, MD,

Department of Gastroenterology, The Affiliated Hospital of Qingdao University, No. 16 Jiangsu Road, Qingdao 266003, China.

Phone: +8617853299218

Email: [lixiaoyu05@163.com](mailto:lixiaoyu05@163.com)



## **Abstract**

### ***Background***

Gastrointestinal stromal tumors (GISTs) and gastrointestinal leiomyomas (GILs) are the most common subepithelial lesions (SELs). All GISTs have malignant potential; however, GILs are considered benign. Current imaging cannot effectively distinguish GISTs from GILs. We aimed to develop an artificial intelligence (AI) system to differentiate these tumors using endoscopic ultrasonography (EUS).

### ***Methods***

The AI system was based on EUS images of patients with histologically-confirmed GISTs or GILs. Participants from four centers were collected to develop and retrospectively evaluate an AI-based system. The system was used when endosonographers considered SELs as GISTs or GILs. We used the system in a multicenter prospective diagnostic test to clinically explore whether joint diagnoses by endosonographers and the AI system can distinguish between GISTs and GILs to improve the total diagnostic accuracy of SELs.

### ***Results***

The AI system was developed using 10439 EUS images from 752 participants with GISTs or GILs. In the prospective test, 132 participants were histologically-diagnosed (36 GISTs, 44 GILs, and 52 other types of SELs) among 508 consecutive subjects. Through joint diagnoses, the total accuracy of endosonographers in diagnosing the 132 histologically-confirmed participants increased from 69.7% (95% confidence interval [CI]: 61.4–76.9%) to 78.8% (95% CI: 71–84.9%;  $p = 0.012$ ). The accuracy of

endosonographers in diagnosing the 80 participants with GISTs or GILs increased from 73.8% (95% CI: 63.1–82.2%) to 88.8% (95% CI: 79.8–94.2%;  $p=0.012$ ).

### ***Conclusions***

We developed an AI-based EUS diagnostic system that can effectively distinguish GISTs from GILs and improve the diagnostic accuracy of SELs.

**Keywords:** deep learning; subepithelial lesion; endoscopic ultrasonography

## Introduction

Subepithelial lesions (SELs) are most commonly found in the stomach (approximately 1 in every 300 endoscopies) [1]. As the most common SELs, gastrointestinal stromal tumors (GISTs) and gastrointestinal leiomyomas (GILs) have much higher incidences than other types of SELs [1]. Most GISTs (60–70%) are located in the stomach, and all GISTs have malignant potential [1–4]. GISTs need to be monitored using endoscopic ultrasonography (EUS) or resected [1,5]. GILs, which are benign tumors, are commonly found in the esophagus and stomach [6]. Resection is only required for GILs with obvious associated symptoms [1]. Considering the differences in prognosis and treatment between GISTs and GILs, accurate differentiation is clinically significant. EUS is considered the most accurate imaging method for evaluating SELs in the gastrointestinal tract [1,7]. However, because of the similar imaging characteristics, it is difficult for endosonographers to distinguish GISTs from GILs using EUS. Additionally, because of the risk of invasive biopsy and uncertain acquisition rate of qualified specimens, the use of fine needle aspiration and “bite-on-bite” pinch biopsies are limited, especially for SELs (<20 mm) [1,8]. Therefore, an accurate non-invasive method for distinguishing GISTs from GILs is warranted.

Artificial intelligence (AI), especially deep learning, has already shown powerful abilities in medical imaging recognition [9-15]. AI can automatically identify quantitative pixel-level features [16,17]. Considering the lack of new breakthroughs in distinguishing between GISTs and GILs in the field of imaging, we wondered whether AI could differentiate these tumors using EUS.



We developed an AI diagnostic system based on preoperative EUS images of histologically-confirmed GISTs and GILs to help endosonographers distinguish these tumors.

## **Methods**

### ***Diagnosis criteria and participants***

This study was comprised of two parts: the development of the AI system and its clinical evaluation (Fig. 1a). Based on the guidelines of the American Society for Gastrointestinal Endoscopy, EUS was performed for screening or pretreatment by endosonographers [1,7]. All participants were derived from consecutive EUS subjects. Based on the World Health Organization's Classification of Tumors, histological diagnosis is considered the reference standard test for SELs [18]. Participants without a definitive pathological diagnosis or report were excluded.

During the development process, EUS images of GISTs or GILs were obtained. Between June 15, 2013 and July 11, 2019, participants with GISTs or GILs from the Affiliated Hospital of Qingdao University Shinan Hospital (AHQSN) were enrolled in the internal dataset to develop and test the AI model. To evaluate whether the AI system applies to different EUS probes, participants with GISTs or GILs from the Affiliated Hospital of Qingdao University Laoshan Hospital (AHQLS), Qilu Hospital QLH, and Qingdao Municipal Hospital East Hospital (QMEH) were enrolled in the external dataset. The models and frequencies of the EUS probes are shown in Supplementary Table 1. Participants from the AHQLS and QLH were enrolled

between November 10 and June 10, 2020. Participants from QMEH were enrolled between January 1, 2018 and June 10, 2020.

During the clinical evaluation process, retrospective and prospective diagnostic tests were performed in three centers (AHQSN, AHQLS, and the Affiliated Hospital of Qingdao University Huangdao Hospital [AHQHD]) to further explore clinical adaptability of the AI system. The estimation of the required sample size of the diagnostic tests is presented in Appendix 1.

In the retrospective diagnostic test, participants with histologically-confirmed GISTs or GILs were obtained from consecutive EUS subjects between June 24 and September 30, 2020. The EUS images, white-light endoscopic images, and medical records were collected to compare the performance of the AI system and endosonographers.

In the prospective diagnostic test, subjects who were found to have gastrointestinal subepithelial protuberant lesions on white-light endoscopy and were ready to undergo EUS were considered as potential prospective participants. Potential participants whose gastrointestinal submucosal bulges were proven to be normal vessels, organ oppression, and extra-gastrointestinal lesions were excluded. Qualified prospective participants were patients with EUS-confirmed SELs. From November 1, 2020 to February 10, 2021, prospective participants were consecutively enrolled and followed up until a clear histological diagnosis was obtained.

### ***Ethics and safety consideration***

This study was approved by the committee of the Affiliated Hospital of Qingdao

University (QYFYWZLL25975). All procedures were performed on the principles of the Declaration of Helsinki. Histological samples were not obtained in this study. Exemption from the need for informed consent from anonymous retrospective data was obtained from the hospitals. The prospective diagnostic test was registered (ChiCTR2000039322) and published in October 2020. Informed consent was obtained from all prospective participants before EUS examinations. All authors had access to the study data and reviewed and approved the final manuscript.

### ***Quality control of EUS images for developing the AI model***

Participants with missing EUS images were excluded. Unqualified EUS images, including low-quality or duplicate images and images with tumors obscured by the measuring lines, were excluded.

For EUS images in the internal and external datasets, we framed and labeled the regions of interest (tumors shown in images) into bounding boxes and extracted them to reduce noise data (Fig. 1b). Using EUS and pathological reports, two endosonographers from AHQSN and AHQLS with >5 years of experience in EUS independently annotated all EUS images. Tumor images with an intersection-over-union score  $\geq 80\%$  were considered qualified. Every final tumor image from different centers was randomly selected from the two qualified tumor images processed by endosonographers. Images with an intersection-over-union score  $< 80\%$  were framed and enrolled after the two endosonographers reached a consensus. The roles of the involved doctors are listed in Supplementary Table 2. All labeled tumor images were resized to 500×500 pixels before model development and evaluation.

### ***Development of the AI model and system***

The AI model was developed using tumor images in the internal dataset, which were randomly assigned to the training set, validation set for AI model development, and internal test set for model evaluation (7:1.5:1.5). The external dataset was used as an external test set for a multicenter test to evaluate the model using different EUS probes.

ResNet-50, as one of the most widely used convolution neuron networks for image recognition, combines performance and running speed [19]. Therefore, the AI model was based on the ResNet-50 with a stochastic gradient descent optimizer. Batch normalization was performed to avoid overfitting. The model was trained on the image level and evaluated at the patient level. The input images of the model were the labeled tumor images. The judgment standards in the training and test processes were the image labels confirmed by histological reports: GIST was a positive label and GIL was a negative label. Considering that receiver operating characteristic curves show a tradeoff between sensitivity and specificity [20], the key model-evaluating metric was the area under the receiver operating characteristic curve (AUC). The model with the best AUC was saved. The other configurations are listed in Supplementary Table 3.

After the AI model was evaluated using the internal and external test sets, it was developed into software using C# programming language.

### ***Design of diagnostic tests***

To further evaluate the AI system, it was compared with human endosonographers

through retrospective and prospective tests. Both diagnostic tests were three-center studies conducted at AHQSN, AHQLS, and AHQHD, which are three of the five branch hospitals of the Affiliated Hospital of Qingdao University Group. Each center has its own endoscopic center, department of pathology, and department of gastrointestinal surgery. These endoscopic centers use the same EUS probes.

In the retrospective diagnostic test, three endosonographers (>10 years of EUS experience) from AHQSN, AHQLS, and AHQHD were invited to diagnose the lesions on EUS images as GISTs or GILs. They were then asked to use the AI system to classify lesions again. White-light endoscopic images, EUS images of participants, and other information such as the size and location of SELs were available to them. The medical histories of the retrospective participants were hidden to the endosonographers to avoid any biases, and endosonographers and retrospective participants were regrouped to ensure that the endosonographers were blind to the participants (Fig. 1c). The researchers in charge of data acquisition were not involved in the diagnostic process.

In the prospective diagnostic test, the AI system was installed on the workstations with a real-time EUS monitor and implemented during EUS examinations. A prospective test was conducted during EUS examinations of the consecutive participants from AHQSN, AHQLS, and AHQHD. First, the endosonographers were asked to specifically classify different types of SELs with EUS on the basis of the guideline [1]. When SELs were considered as GISTs or GILs, the AI system would be then used to jointly diagnose them. Every AI diagnosis was based on at least five

tumor images directly captured from real-time EUS monitors, which clearly showed the SELs and layers of the gastrointestinal tract. Joint diagnoses of the participants were based on AI diagnoses. The predictions of AI diagnoses were blind to the prospective participants to avoid interference with their clinical process.

### ***Statistical analysis***

The metrics, including diagnostic accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), were evaluated using the formulae listed in Supplementary Table 4. The 95% confidence intervals (CIs) were calculated using the modified Wald method. Mean pixel values were compared using T test. Diagnostic accuracy, sensitivity, and specificity were compared using chi-squared test or McNemar's test; PPV and NPV were compared using marginal logistic regression [21]. All statistical tests were two-sided, with a significance level of at least 0.05. Python (version 3.7.0) and SPSS 22.0 (IBM, CA, USA) were used for statistical analyses.

## **Results**

### ***Dataset for model development and performance evaluation***

During model development, 10,439 images from 752 participants with histologically-confirmed GISTs or GILs were used to train and retrospectively evaluate the AI model (Table 1, Supplementary Table 5, and Appendix 2).

From June 15, 2013 to July 11, 2019, 702 participants with GISTs and GILs from AHQSN were recruited in the internal dataset. A total of 227 participants with GISTs



and 266 participants with GILs were randomly assigned to the training set. The others were randomly assigned to the validation (47 GISTs and 58 GILs) and internal test (50 GISTs and 54 GILs) sets.

In the external centers, 24 participants with GISTs and 26 participants with GILs were enrolled in the external test set. In detail, there were 23 cases from AHQLS, 20 cases from QLH, and 7 cases from QMEH.

### ***Performance of the AI model***

The AUC of the AI model was 0.986 (95% CI: 0.967–1) for the internal test set and 0.642 (95% CI: 0.466–0.801) for the external test set (Fig. 2a, b). The predictive scores ranged from 0 to 1. According to the predicted results of the internal test set, the model reached the highest diagnostic accuracy when the cutoff score for diagnosis was 0.59 (GIST <0.59; GIL ≥0.59). Therefore, the cutoff value was set to 0.59. The diagnostic accuracy of the AI model in the internal test set was 96.2% (95% CI: 90.2–98.8%). The sensitivity, specificity, PPV, and NPV were 98% (95% CI: 88.5–100%), 94.4% (95% CI: 84.3–98.7%), 94.2% (95% CI: 83.8–98.6%) and 98.1% (95% CI: 88.9–100%), respectively (Fig. 2a).

The diagnostic accuracy of the model in the external test set was 66.0% (95% CI: 52.1–77.6%). The corresponding sensitivity, specificity, PPV and NPV were 45.8% (95% CI: 27.9–64.9%), 84.6% (95% CI: 65.9–94.5%), 73.3% (95% CI: 47.6–89.5%) and 62.9% (95% CI: 46.3–76.9%). The diagnostic accuracy of the AI system was 91.3% (95% CI: 72–98.8%) for AHQLS participants, 50.0% (95% CI: 30–70%) for QLH participants, and 28.6% (95% CI: 7.6–64.8%) for QMEH participants.

Therefore, the performance of the AI model in AHQSN and AHQLS, which use the same EUS probes, may be accurate. In contrast, the accuracies of the AI model in QLH and QMEH, which use different EUS probes from AHQSN, were low. More details on the internal and external tests are listed in Supplementary Table 6.

The EUS images from the four centers are shown in Supplementary Figure 1. Endosonographers can observe the differences in image quality. To quantify the difference, we randomly selected 4430 tumor images from the four centers to analyze their image pixel values (Supplementary Table 7). The mean pixel values from AHQSN ( $50.7 \pm 31.3$ ) and AHQLS ( $51.6 \pm 30.2$ ) were not significantly different from each other ( $p = 0.97$ ), but they were significantly different from the pixel values from QLH ( $61.0 \pm 35.4$ ,  $p = 0.021$ ) and QMEH ( $17.3 \pm 19.4$ ,  $p < 0.0001$ ), suggesting differences in imaging between different devices.

After evaluating the AI model, it was developed into an AI system with a graphical user interface (Fig. 2c and Video 1). The AI system is used for EUS (UM-DP12-25R, 20MHz; UM-DP20-25R [Olympus, Tokyo, Japan]) without being adjusted for different EUS operating frequencies. Other information on the AI system is provided in Appendix 3.

### ***Results of the diagnostic tests***

During the clinical evaluation of the AI system, we conducted two diagnostic tests. For the retrospective diagnostic test, 84 participants with histologically-confirmed GISTs or GILs were recruited from 532 consecutive subjects with SELs diagnosed on EUS (Table 2). The accuracy of the three endosonographers in distinguishing GISTs

and GILs was 70.2% (95% CI: 59.7–79%) (Fig. 2d). The accuracy of the AI system was significantly higher than that of the endosonographers at 94% (95% CI: 86.5–97.8%,  $p=0.00033$ ). The sensitivity, specificity, PPV and NPV of the AI system were greater than those of the endosonographers (90% [95% CI: 73.6–97.3%] vs. 60% [95% CI: 42.3–75.4%],  $p=0.035$ ; 96.3% [95% CI: 86.7–99.7%] vs. 75.9% [95% CI: 62.9–85.5%],  $p=0.003$ ; 93.1% [95% CI: 77–99.2%] vs. 58.1% [95% CI: 40.7–73.6%],  $p=0.00019$ ; 94.5% [95% CI: 84.6–98.7%] vs. 77.4% [95% CI: 64.3–86.7%],  $p=0.041$ ).

The processes and results of the prospective diagnostic tests are shown in Fig. 3 and Table 3. A total of 508 prospective consecutive subjects were included in the test. Among the 508 participants, 256 participants were diagnosed with GISTs or GILs by endosonographers and received AI diagnosis; the other 252 participants were considered to have other kinds of SELs (for example: neuroendocrine tumor and ectopic pancreas) and did not received AI diagnosis. Eighty-one of the 256 participants who were diagnosed with GISTs or GILs by endosonographers were histologically diagnosed (36 GISTs, 41 GILs and 4 other types of SELs). In the 252 participants who were considered to have other types of SELs by endosonographers, there are 56 of them with histological examinations (3 GILs, 48 other kinds of SELs and 5 without clear diagnosis). Totally, 132 of the 508 participants obtained clear histological diagnoses (36 GISTs, 44 GILs and 52 other types of SELs).

When combined with the AI system, the total diagnostic accuracies of the endosonographers in diagnosing the 132 histologically-confirmed prospective

participants increased from 69.7% (95% CI: 61.4–76.9%) to 78.8% (95% CI: 71–84.9%;  $p = 0.012$ ). The accuracy of the endosonographers in distinguishing the 36 participants with GISTs from the 44 participants with GILs increased from 73.8% (95% CI: 63.1–82.2%) to 88.8% (95% CI: 79.8–94.2%;  $p = 0.012$ ). Among the 36 participants with GISTs and 44 participants with GILs, 77 of them were initially diagnosed as GISTs or GILs by endosonographers and received AI diagnoses. Among the 77 participants, the accuracy, specificity, and PPV of joint diagnoses were 92.2% (95% CI: 83.7–96.7%, Fig. 2e), 95.1% (95% CI: 83–99.5%), and 94.1% (95% CI: 80–99.4%), which were greater than those of independent diagnoses (76.6% [95% CI: 66–84.8%,  $p = 0.012$ ], 65.9% [95% CI: 50.5–78.5%,  $p = 0.002$ ], and 69.6% [95% CI: 55.1–81%,  $p = 0.0004$ ], respectively). However, the sensitivity and NPV of joint diagnoses were not significantly higher than those of independent diagnoses (88.9% [95% CI: 74.1–96.2%] vs. 88.9% [95% CI: 74.1–96.2%],  $p = 1$ ; 90.7% [95% CI: 77.8–96.9%] vs. 87.1% [95% CI: 70.5–95.5%],  $p = 0.74$ ). More details on diagnostic tests are shown in Supplementary Tables 5, 6, and 8.

## Discussion

In this study, we introduced an AI system for distinguishing GISTs from GILs and shared its graphical software. Multicenter data sets, including 752 participants with histologically-confirmed GISTs or GILs, were used to develop and evaluate the AI model. Moreover, we conducted both a retrospective and prospective diagnostic test in three centers to clinically evaluate the AI system. Although the current consensus is

that there are no definitively detectable imaging features to guide human endosonographers in differentiating GISTs from GILs, the AI system showed good performance in both diagnostic tests [1]. The combination of endosonographers and the AI system significantly improved the diagnostic accuracy of SELs using EUS. The AI system is expected to decrease the misdiagnosis rate of GISTs and GILs, which will likely help patients avoid unnecessary EUS, invasive biopsies, and surgeries.

According to published guidelines, stacked or “bite-on-bite” pinch biopsies may be attempted in SELs; however, the accuracy is often low [1,22,23]. The acquisition of qualified specimens for SELs <20 mm is more difficult, which may lead to low diagnostic accuracy [8,24,25]. Therefore, the guidelines for EUS-guided sampling by the European Society of Gastrointestinal Endoscopy does not recommend performing EUS-FNA in patients with SELs <20 mm [8]. Conversely, the AI system can significantly improve the diagnostic accuracy of SELs without the limitation of SEL size, avoiding invasive biopsy. All tumor images were resized to 500×500 pixels to reduce the influence of tumor size on AI diagnosis. According to the prospective diagnostic test, the diagnostic accuracy of the AI system for GISTs and GILs <20 mm in size was close to that for GISTs and GILs ≥20 mm in size (90% [95% CI: 78.2–96.1%] vs. 96.3% [95% CI: 80.2–100%],  $p=0.31$ ; Supplementary Table 6).

Previous studies on the application of AI in medical imaging have usually focused on human-distinguishable issues, such as gastrointestinal cancer detection, delineating margins of early gastric cancer, and grading of cancer [9,26,27]. Conversely, the present study focused on GISTs and GILs, which are generally

considered indistinguishable in imaging examinations. The AI system can effectively distinguish between GISTs and GILs, suggesting the great potential of AI in solving problems that human medical experts have difficulty solving.

Additionally, most previously reported AI models require expensive backend servers and complex technical support [9,28]. To improve the practicality of the AI system, we developed a compatible system and graphical user interface for the AI model based on the ResNet-50. The ResNet-50 combines performance and running speed, which allows the AI system to achieve high accuracy and work on entry-level computers with a Windows 64-bit system [19]. We believe that hardware compatibility and ease of evaluation are important, especially in the initial stages of exploring whether AI can help endosonographers distinguish between GISTs and GILs. For hospitals, the system does not incur extra costs for equipment updates. As the AI system can improve diagnostic accuracy without significantly increasing the examination fees, it is more helpful for patients from developing regions. Furthermore, the system does not require additional training for operators. The entire diagnostic process, including screenshot and selection of clear EUS images, framing of tumor images, and data analysis takes about one minute, which can avoid the waiting time for biopsy reports, thus enabling endoscopists to quickly make clinical decisions and optimize the treatment plans.

Although the AI system showed high performance, it also had some limitations. First, this AI model was proven effective only for specific EUS probes. The pixel value analysis of the EUS images shows that different EUS probes have imaging



differences, which probably affect AI performance. The AI model may overfit the ultrasonographic images taken by the specific devices. However, this study may still be helpful in developing models for different EUS probes by transfer learning and federated learning, which can decrease the amount of data required and avoid concerns about patient privacy. Second, the AI system was designed to distinguish GISTs and GILs and does not work on other types of SELs. Therefore, the use of AI systems still relies on the experience of endosonographers. We have attempted to enroll other types of SELs, such as Schwannomas and glomus tumors, into the classifier development; however, because their incidence is very low and overfitting was unavoidable without sufficient data, we could not train a reliable model. Larger multicenter or international studies are expected to solve these two problems. Third, approximately 30% of consecutive EUS subjects with SELs obtained a clear histological diagnosis, which caused verification bias. Verification bias is common in diagnostic tests, as it may not be ethical, practical, or cost-effective to obtain a histological diagnosis in every patient [29,30]. With consecutive subjects diagnosed with GISTs or GILs on EUS in AHQSN between June 24, 2020 and September 30, 2020, as an example, the subjects diagnosed with GIST on EUS were more likely to obtain a histological diagnosis (Supplementary Table 9 and Appendix 4). Therefore, the sensitivities of endosonographers and the AI system may be overestimated, and the specificities may be underestimated. Detailed analyses of verification bias are provided in Appendix 4.

In summary, we have developed an AI-based diagnostic system using EUS

images that can differentiate GISTs from GILs. This system can improve the diagnostic accuracy of SELs.

### **Data sharing**

The document regarding the AI system will be shared on (the files will be available after publication).

### **Competing interests**

All authors declare that they have no conflict of interest.

### **Acknowledgments**

We thank Dr. L. Zhang from the Department of Computer Science and Technology at Fudan University for his help with computer programming, and Dr. S. Zhang from Qilu Hospital for his help with data collection. Meanwhile, we thank Dr. Yueping Jiang, Dr. Junying Tan, Dr. Man Xie, Dr. Xiaowei Wang, Dr. Xueli Ding, Dr. Xue Jing, Dr. Hui Ju, Dr. Fuguo Liu, Dr. Qingdong Liu, Dr. Wei Zhang and Dr. Hao Yuan from the Affiliated Hospital of Qingdao University for their help in diagnostic tests. This work was supported by the National Natural Science Foundation of China (Grant81802777) and “Clinical medicine + X” scientific research project of Affiliated Hospital of Qingdao University.

### **References**

1. Standards of Practice C, Faulx AL, Kothari S et al. The role of endoscopy in

subepithelial lesions of the GI tract. *Gastrointest Endosc* 2017; 85: 1117-1132

2. Miettinen M, Lasota J. Gastrointestinal stromal tumors: Pathology and prognosis at different sites. *Semin Diagn Pathol* 2006; 23: 70-83
3. Chandrasekhara V, Ginsberg GG. Endoscopic management of gastrointestinal stromal tumors. *Curr Gastroenterol Rep* 2011; 13: 532-539
4. Nishida T, Blay JY, Hirota S et al. The standard diagnosis, treatment, and follow-up of gastrointestinal stromal tumors based on guidelines. *Gastric Cancer* 2016; 19: 3-14
5. Zhou P, Zhong Y, Li Q. Chinese Consensus on Endoscopic Diagnosis and Management of Gastrointestinal Submucosal Tumor (Version 2018). *Chinese Journal of Practical Surgery* 2018; 21: 841-852
6. Nishida T, Kawai N, Yamaguchi S et al. Submucosal tumors: comprehensive guide for the diagnosis and therapy of gastrointestinal submucosal tumors. *Dig Endosc* 2013; 25: 479-489
7. Landi B, Palazzo L. The role of endosonography in submucosal tumours. *Best Pract Res Clin Gastroenterol* 2009; 23: 679-701
8. Dumonceau JM, Deprez PH, Jenssen C et al. Indications, results, and clinical impact of endoscopic ultrasound (EUS)-guided sampling in gastroenterology: European Society of Gastrointestinal Endoscopy (ESGE) Clinical Guideline - Updated January 2017. *Endoscopy* 2017; 49: 695-714
9. Luo H, Xu G, Li C et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study.

Lancet Oncol 2019; 20: 1645-1654

10. Byrne MF, Chapados N, Soudan F et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut 2019; 68: 94-100
11. Kather JN, Pearson AT, Halama N et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 2019; 25: 1054-1056
12. Bi WL, Hosny A, Schabath MB et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin 2019; 69: 127-157
13. Song Z, Zou S, Zhou W et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nat Commun 2020; 11: 4294
14. Ikenoyama Y, Yoshio T, Tokura J et al. Artificial intelligence diagnostic system predicts multiple Lugol-voiding lesions in the esophagus and patients at high risk for esophageal squamous cell carcinoma. Endoscopy 2021, DOI: 10.1055/a-1334-4053.
15. Wu L, He X, Liu M et al. Evaluating the Effects of An Artificial Intelligence System on Endoscopy Quality and Preliminarily Testing its Performance on Detecting Early Gastric Cancer: A Randomized Controlled Trial. Endoscopy 2021, DOI: 10.1055/a-1350-5583.
16. Trister AD, Buist DSM, Lee CI. Will Machine Learning Tip the Balance in Breast Cancer Screening? JAMA Oncol 2017; 3: 1463-1464
17. Ding Z, Shi H, Zhang H et al. Gastroenterologist-Level Identification of Small-Bowel Diseases and Normal Variants by Capsule Endoscopy Using a Deep-Learning

Model. *Gastroenterology* 2019; 157: 1044-1054 e1045

18. Nagtegaal ID, Odze RD, Klimstra D et al. The 2019 WHO classification of tumours of the digestive system. *Histopathology* 2020; 76: 182-188

19. He K, Zhang X, Ren S et al. Deep Residual Learning for Image Recognition. In, *IEEE Conference on Computer Vision & Pattern Recognition*; 2016

20. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. *J Biomed Inform* 2014; 48: 193-204

21. Leisenring W, Alonzo T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics* 2000; 56: 345-351

22. Hunt GC, Smith PP, Faigel DO. Yield of tissue sampling for submucosal lesions evaluated by EUS. *Gastrointest Endosc* 2003; 57: 68-72

23. Menon L, Buscaglia JM. Endoscopic approach to subepithelial lesions. *Therap Adv Gastroenterol* 2014; 7: 123-130

24. Antonini F, Delconte G, Fuccio L et al. EUS-guided tissue sampling with a 20-gauge core biopsy needle for the characterization of gastrointestinal subepithelial lesions: A multicenter study. *Endosc Ultrasound* 2019; 8: 105-110

25. Akahoshi K, Oya M, Koga T et al. Clinical usefulness of endoscopic ultrasound-guided fine needle aspiration for gastric subepithelial lesions smaller than 2 cm. *J Gastrointest Liver Dis* 2014; 23: 405-412

26. Nagpal K, Foote D, Tan F et al. Development and Validation of a Deep Learning Algorithm for Gleason Grading of Prostate Cancer from Biopsy Specimens. *JAMA Oncol* 2020, DOI: 10.1001/jamaoncol.2020.2485.

27. Ling T, Wu L, Fu Y et al. A Deep Learning-based System for Identifying

Differentiation Status and Delineating Margins of Early Gastric Cancer in Magnifying Narrow-band Imaging Endoscopy. *Endoscopy* 2020, DOI: 10.1055/a-1229-0920.

28. Gong D, Wu L, Zhang J et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. *Lancet Gastroenterol Hepatol* 2020; 5: 352-361

29. O'Sullivan JW, Banerjee A, Heneghan C et al. Verification bias. *BMJ Evid Based Med* 2018; 23: 54-55

30. Schmidt RL, Jedrzkiewicz JD, Allred RJ et al. Verification bias in diagnostic accuracy studies for fine- and core needle biopsy of salivary gland lesions in otolaryngology journals: a systematic review and analysis. *Head Neck* 2014; 36: 1654-1661



## Figure legends

### Fig. 1 Study framework.

- a. Flowchart of AI system development, and clinical evaluation.
- b. Labeling and sizing of tumor images. Regions of interest (tumor images) in green or red bounding boxes were labeled, extracted, and resized to 500×500 pixels.
- c. Grouping of the participants in retrospective diagnostic test and diagnostic groups from different endoscopic centers

AUC = area under the receiver operating characteristic curve;

GIST = gastrointestinal stromal tumor, GIL = gastrointestinal leiomyoma;

AHQHD = the Affiliated Hospital of Qingdao University Huangdao Hospital;

AHQSN = the Affiliated Hospital of Qingdao University Shinan Hospital;

AHQLS = the Affiliated Hospital of Qingdao University Laoshan Hospital.

### Fig. 2 Performance of the AI model and results of the diagnostic tests.

- a. The performance of the AI model in the internal and external test sets.
- b. AI model prediction receiver operating characteristic curves for the internal and external test sets.
- c. Interface of the AI system.
- d. Retrospective diagnostic test result.

- e. Diagnostic test result of prospective participants who were diagnosed as GISTs or GILs through both EUS and histological examination.

PPV= positive predictive value, NPV=negative predictive value.

**Fig. 3 Flowchart of the prospective diagnostic test.**

Reason A: The gastrointestinal protuberant lesions in 156 potential participants were diagnosed as normal vessels, organ oppression and extra-gastrointestinal lesions. Four potential participants rejected the diagnostic tests.

Reason B: These participants rejected invasive examinations or treatment, and chose to undergo regular medical examinations, or go to other centers for treatment.

**Video image**

**Video legends**

**Video 1. How to use the AI system.**

Step 1. Click and open the AI system software.

Step 2. Click the button “Screenshot”, and frame the tumor shown in the EUS images or videos.

Step 3. Click the button “Automatic diagnosis”. It takes several seconds.

Step 4. The system automatically gives a diagnosis in few seconds.

## Supplementary material

### Authors:

Xintian Yang, Han Wang, Qian Dong, Yonghong Xu, Hua Liu, Xiaoying Ma, Jing Yan, Qian Li, Chenyu Yang, Xiaoyu Li

**Paper Title: An artificial intelligence system for distinguishing between gastrointestinal stromal tumors and leiomyomas using endoscopic ultrasonography (with video)**

### Appendix 1. Estimation of required sample size in the prospective diagnostic test.

The total sample sizes based on sensitivity and specificity respectively are:

$$N = \frac{Z_{\frac{\alpha}{2}}^2 \hat{P}(1-\hat{P})}{d^2 \times Prev}$$

$N$  = total required sample size;  $\alpha = 0.05$ ,  $Z_{\frac{\alpha}{2}} = 1.96$ ;  $\hat{P}$  = predetermined value of sensitivity (or specificity);  $d$  = the maximum marginal error of estimate;  $Prev$  = prevalence of disease (gastrointestinal stromal tumor [GIST] = positive label vs. gastrointestinal leiomyoma [GIL] = negative label)[1].

In the prospective diagnostic test, predetermined values of sensitivity and specificity ( $\hat{P}$ ) are 0.90 and 0.90. The prevalence of GIST and GIL ( $Prev$ ) are 0.46 and 0.54 on the basis of the proportion of 702 cases of GISTs and GILs. In order the maximum marginal error of estimate ( $d$ ) does not exceed from 10% with 95% confidence level ( $1-\alpha$ ), the total required sample size can be driven by plugging the

above values in the prospective diagnostic test as follows:

$$N = \frac{1.96^2 \times 0.90(1-0.90)}{0.1^2 \times 0.46} \approx 76; N = \frac{1.96^2 \times 0.90(1-0.90)}{0.1^2 \times 0.54} \approx 65$$

The larger  $N$  ( $N=76$ ) was chosen as the required sample size in the prospective test set. More than 76 cases of patients with histologically-confirmed GISTs or GILs should be enrolled in the diagnostic test.

## **Appendix 2. Assignment of the participants and tumor images in internal dataset and external dataset.**

According to the data control criteria, 702 respective participants with GISTs and GILs were recruited in the internal dataset. Annotated tumor images from AHQSN were 9926, including 4595 tumor images from 324 participants with GISTs and 5331 tumor images from 378 participants with GILs. A total of 227 participants with GISTs and 3299 tumor images and 266 participants with GILs and 3831 tumor images were randomly categorized into the training set. The other cases in the AHQSN retrospective cohort were randomly assigned to the validation set (642 tumor images from 47 participants with GISTs, 820 tumor images from 58 participants with GILs) and the internal test set (654 tumor images from 50 participants with GISTs, 680 tumor images from 54 participants with GILs).

In the external dataset, 294 tumor images from 24 participants with GISTs and 219 tumor images from 26 participants with GILs were enrolled in the external test set (Table 1). In detail, there were 23 cases with 343 tumor images from AHQLS, 20

cases with 88 tumor images from QLH, and 7 cases with 82 tumor images from QMEH. All of the external participants and tumor images were assigned to external test set.

### **Appendix 3. Information of the AI system with a graphical user interface.**

The AI system uses the central processing unit (CPU) version PyTorch frameworks. In AHQHD, AHQSN and AHQLS, the model ran on workstations with AMD Ryzen 7 2700 CPU (AMD, CA, USA), Intel Core i5-4590 CPU, or i3-8100 (Intel, CA, USA). The AI system can distinguish between GIST and GIL using still endoscopic ultrasonography (EUS) images, stored EUS videos or real-time EUS videos. While endoscopists are performing EUS, other endosonographers can use the AI system on the workstation with a real-time endoscopic monitor. The operator frames the tumor shown in the EUS images or videos to input tumor images into the AI system, and the prediction results are presented on the interface within 10 seconds thereafter (Video 1). Generally speaking, an EUS examination takes 20-40 minutes. The entire process of AI diagnosis, including tumor image capture and data analysis, takes about 1 minute and does not affect the endosonographers. We tested the running time of the AI system on workstations with different CPUs. The results were listed in Supplementary table 12. In addition to CPUs, memories, disks and the size of tumor images can also influence its running speed.

#### **Appendix 4. Details of verification bias analysis.**

##### ***1. Details of the enrollment and exclusion of the internal dataset, retrospective diagnostic test set and prospective test set.***

The internal dataset was used in developing the AI model. Totally, 5766 consecutive subjects were diagnosed as subepithelial lesions (SELs) on EUS between June 15, 2013 and July 11, 2019 in AHQSN. There were 2264 patients diagnosed with GISTs or GILs on EUS, 699 (30.9%) of whom obtained histological diagnosis (Supplementary table 9). Among the 699 patients, there are 656 cases of histologically confirmed GISTs or GILs, 38 cases of other kinds of histologically confirmed SELs, and 5 cases with unclear histological diagnosis (Supplementary Table 9 and 10). In addition to the participants diagnosed with GISTs or GILs on EUS, there were 54 cases with histologically confirmed GISTs or GILs who had been wrongly diagnosed with other kinds of SELs, such as neuroendocrine tumors, schwannoma and glomus tumors, before the histological examinations (Supplementary table 11). According to the asks of data control (shown in the Method part in the paper), 8 cases were excluded because of data missing or unclear images. In total, there are 702 cases with GISTs or GILs included the development of the AI system.

Similarly, we also enrolled participants from consecutive EUS subjects into the retrospective diagnostic test set and prospective diagnostic test set. Approximately



30% of these subjects obtained histological diagnosis. To avoid selection bias, the included participants in the study were derived from consecutive patients. Only 1.1% (8/710) histologically-confirmed patients with GISTs or GILs were excluded because of data missing or unclear images. Compared with the total number of included participants, we believed that the selection bias was limited. Meanwhile, the good performance of the AI system in the prospective diagnostic test also shows that selection bias may not be prominent.

## **2. Analysis of verification bias.**

If we take all the subjects in account, whether they obtain a histological diagnosis or not, there were about 70% patients without histological diagnosis in the process of collecting participants in the study (Supplementary Table 8, 9). Partial verification bias is caused when not all patients were confirmedly diagnosed by the same reference test (gold standard) [2]. Verification bias is very common in diagnostic test, for it may not be ethical, practical or cost-effective to obtain a histological diagnosis (reference standard test) in every patient [3]. Previous studies show that 84% (81/95) studies had verification bias[3]. Only 5 studies of 81 studies provided sufficient information to correct the verification bias. Two studies provided a flow diagram. We have already provided a flow diagram and details of subjects. Herein, we provided some analyses about the verification bias in the study.

In this study, we compared and discussed the performance of endosonographers

and the AI system. In the consecutive subjects diagnosed with GISTs or GILs during June 15, 2013 and July 11, 2019 in AHQSN, the ratio of reference standard test in the subjects diagnosed as GILs on EUS by endosonographers were 24.1% (298/1238). The ratio of reference standard test in the subjects diagnosed as GISTs on EUS by endosonographers were 39.1% (401/1026). The subjects diagnosed with GISTs on EUS tend to undergo reference standard test ( $p < 0.0001$ ). Therefore, the sensitivities and specificities of the endosonographers and the AI system were overestimated and the specificities were underestimated.

To modify the verification in the two competing diagnostic tests, some researchers provided with some methods to calculate modified sensitivities or specificities under some assumptions:

Assumption1. Missing completely at random (MCAR)

MCAR means that missing data (cases without reference standard test) are not correlated to any results of diagnostic tests or any variables (such as sex, age and symptoms).

Andrzej S.Kosinski and Huiman X.Barnhart reported a method to modify sensitivity and specificity under Assumption 1 [4, 5].

	T1 = 1		T2 = 2	
	T2 = 1	T2 = 0	T2 = 1	T2 = 0
V = 1				
D = 1	a <sub>11</sub>	a <sub>10</sub>	a <sub>01</sub>	a <sub>00</sub>
D = 0	b <sub>11</sub>	b <sub>10</sub>	b <sub>01</sub>	b <sub>00</sub>
V = 0	u <sub>11</sub>	u <sub>10</sub>	u <sub>01</sub>	u <sub>00</sub>

Total	m <sub>11</sub>	m <sub>10</sub>	m <sub>01</sub>	m <sub>00</sub>
-------	-----------------	-----------------	-----------------	-----------------

T1 = Diagnostic test 1, T2 = Diagnostic test 2, V = verification (1 = positive, 0 = negative), D = status of disease (1 = disease, 0 = non-disease),

The sensitivity(se) and specificity(sp) of the two diagnostic tests were as follow:

$$se_1 = \frac{a_{11} + a_{10}}{a_{11} + a_{10} + a_{01} + a_{00}}, sp_1 = \frac{b_{01} + b_{00}}{b_{11} + b_{10} + b_{01} + b_{00}}$$

$$se_2 = \frac{a_{11} + a_{01}}{a_{11} + a_{10} + a_{01} + a_{00}}, sp_2 = \frac{b_{10} + b_{00}}{b_{11} + b_{10} + b_{01} + b_{00}}$$

Assumption2. Missing at random (MAR)

MAR means that missing data are only correlated to the diagnostic test results, but not correlated to other variables (such as sex, age and symptoms)

Zhou reported a maximum likelihood method to modify verification bias under Assumption 2:

$$se_1 = \sum_{t=0}^1 \frac{a_{1t}}{a_{1t}+b_{1t}} * m_{1t} \bigg/ \sum_{i=0}^1 \sum_{t=0}^1 \frac{a_{it}}{a_{it}+b_{it}} * m_{it}, \quad sp_1 = \sum_{t=0}^1 \frac{b_{0t}}{a_{0t}+b_{0t}} * m_{0t} \bigg/ \sum_{i=0}^1 \sum_{t=0}^1 \frac{b_{it}}{a_{it}+b_{it}} * m_{it}$$

$$se_2 = \sum_{i=0}^1 \frac{a_{i1}}{a_{i1}+b_{i1}} * m_{i1} \bigg/ \sum_{i=0}^1 \sum_{t=0}^1 \frac{a_{it}}{a_{it}+b_{it}} * m_{it}, \quad sp_2 = \sum_{i=0}^1 \frac{b_{i0}}{a_{i0}+b_{i0}} * m_{i0} \bigg/ \sum_{i=0}^1 \sum_{t=0}^1 \frac{b_{it}}{a_{it}+b_{it}} * m_{it}$$

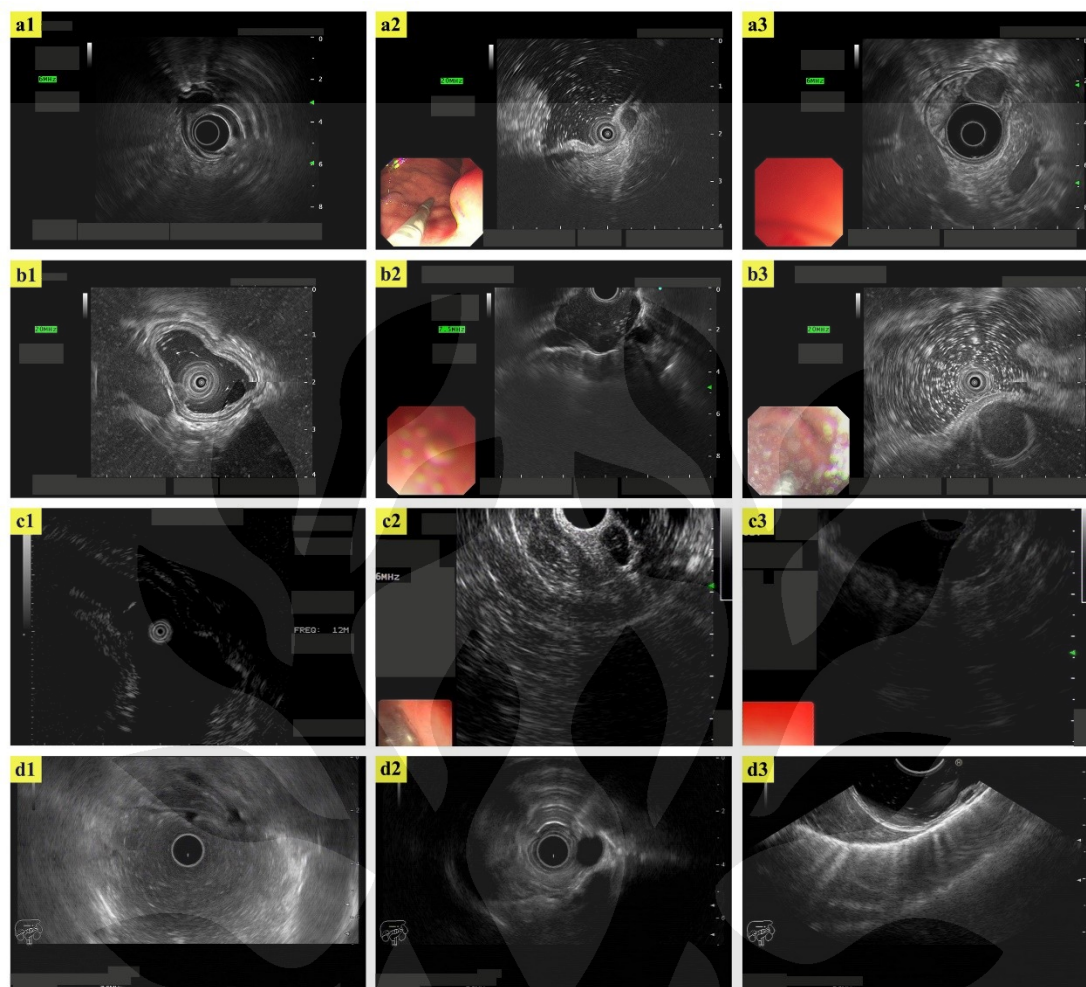
$$t = T1, i = T2$$

Zhou's method is one of the most widely used method to modify verification in a two-phase diagnostic test [3, 6]. There are some other Bayes methods or likelihood methods to modify verification bias in multiple diagnostic tests [7, 8]. But all of them rest on assumptions that are difficult to verify.

In the present study, because of the evidence and recommendation in the guideline, lesions > 20mm in size tend obtain histological diagnosis. As Supplementary Figure 2 shows, patients with larger lesions have bigger ratio of histological diagnosis (<10mm: 15.9% [220/1385] vs. ≥10mm, <20mm: 46.9% [246/525] vs. ≥20mm: 66.1% [232/351];  $p < 0.0001$ ), which is in accord with the guideline and our clinical experience. Moreover, the subjects >60 years has smaller ratio of histological examination than the subjects <60 years (26.9% [230/855] vs. 33.3% [469/1409],  $p = 0.0014$ ). Since older subjects are more likely to suffer from cardiovascular disease or other chronic diseases, subjects and endoscopists tend to

avoid invasive examination and treatment. In addition, according to recommendations of the guideline or our clinical experience, the locations of lesions, symptoms and income of the subjects also affect histological test [9]. Especially, the influence of the income of patients is very complex. In general, verification of subjects is influenced by multivariable. As a result, the methods using under Assumption 1 and Assumption 2 are not applicable to this study. We did not find suitable method of quantitatively calculating modified sensitivities and specificities in the study.

### **Supplementary Figure 1**



### Supplementary Figure 1 legend

#### EUS images by different devices

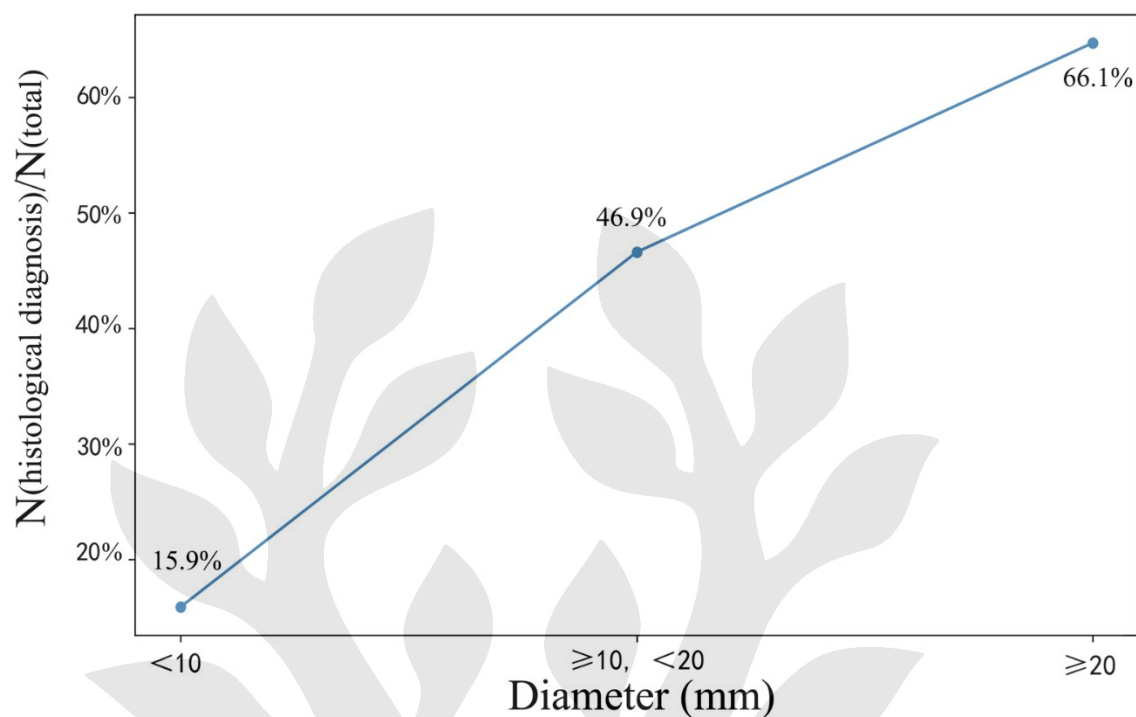
a1-a3: from AHQSH, by UM-DP12-25R, UM-DP20-25R Olympus, Tokyo, Japan

b1-b3: from AHQLS, by UM-DP12-25R, UM-DP20-25R Olympus, Tokyo, Japan

c1-c3: from QMEH, by UM-2R, UM-3R Olympus, Tokyo, Japan

d1-d3: from QLH, by EG-3670URK, EG-3870UTK Pentax, Tokyo, Japan

### Supplementary Figure. 2



#### Supplementary Figure 2 legend

The ratio of participants with histological diagnosis to consecutive subjects diagnosed as GISTs or GILs on EUS in the process of collecting the internal dataset.

GIST = gastrointestinal stromal tumor, GIL = gastrointestinal leiomyoma, EUS = endoscopic ultrasonography.

**Supplementary Table 1: EUS probes and corresponding frequency in different centers.**

	<b>AHQSN</b>	<b>AHQLS</b>	<b>QLH</b>	<b>QMEH</b>	<b>AHQHD</b>
<b>EUS probes</b>	UM-DP12-25R, 20MHz:UM-DP20- 25R (Olympus, Tokyo, Japan)	UM-DP12-25R, 20MHz:UM- DP20-25R (Olympus, Tokyo, Japan)	EG-3670URK, EG-3870UTK (Pentax, Tokyo, Japan)	UM-2R, UM- 3R (Olympus, Tokyo, Japan)	UM-DP12-25R, 20MHz:UM- DP20-25R (Olympus, Tokyo, Japan)
<b>Frequency (MHz)</b>	6,7.5, 10, 12 and 20	6,7.5, 10, 12 and 20	5, 7.5, and 10	6, 12, and 20	6,7.5, 10, 12 and 20

**AHQSN** = Affiliated Hospital of Qingdao University Shinan Hospital;

**AHQLS** = Affiliated Hospital of Qingdao University Laoshan Hospital;

**QLH** = Qilu Hospital;

**QMEH** = Qingdao Municipal Hospital East Hospital;

**AHQHD** = Affiliated Hospital of Qingdao University Huangdao Hospital.

**Supplementary Table 2. Roles of doctors in the data processing and diagnostic tests.**



Name	Position	Education	Experience	Contribution	Hospital
Xiaoyu Li	Deputy Chief doctor	M.D.	10 years	Annotation and quality control	AHQSN
Qian Li	Attending doctor	M.D.	5 years	Annotation and quality control	AHQLS
Yonghong Xu	Chief doctor	M.D.	15 years	Prospective diagnostic tests	AHQSN
Hua Liu	Deputy Chief doctor	M.D.	11 years	Retrospective and prospective diagnostic tests	AHQSN
Tan Junying	Attending doctor	M.D.	10 years	Retrospective and prospective diagnostic tests	AHQHD
Yueping Jiang	Chief doctor	M.D.	15 years	Retrospective and prospective diagnostic tests	AHQLS
Man Xie	Deputy Chief doctor	M.D.	8 years	Prospective diagnostic test	AHQLS
Xiaowei Wang	Deputy Chief doctor	M.D.	6 years	Prospective diagnostic test	AHQSN
Xueli Ding	Deputy Chief doctor	M.D.	12 years	Prospective diagnostic test	AHQSN
Xue Jing	Deputy Chief doctor	M.D.	10 years	Prospective diagnostic test	AHQLS
Hui Ju	Attending doctor	M.D.	14 years	Prospective diagnostic test	AHQHD
Hao Yuan	Chief doctor	M.D.	16 years	Prospective diagnostic tests	AHQHD
Fuguo Liu	Deputy Chief doctor	M.D.	14 years	Prospective diagnostic test	AHQHD

Junheng Liu	Resident doctor	M.D.	3 years	Prospective diagnostic test	AHQHD
Qingdong Mao	Attending doctor	M.D.	8 years	Prospective diagnostic test	AHQHD
Wei Zhang	Deputy Chief doctor	M.D.	9 years	Prospective diagnostic test	AHQHD
Xintian Yang	Trainee doctor	M.M.	3 years	Data arrangement	AHQSN
Jing Yan	Trainee doctor	M.M.	1 year	Data arrangement	AHQSN

**AHQSN** = Affiliated Hospital of Qingdao University Shinan Hospital;

**AHQLS** = Affiliated Hospital of Qingdao University Laoshan Hospital;

**AHQHD** = Affiliated Hospital of Qingdao University Laoshan Hospital.

**Supplementary Table 3. Configurations of the AI model.**

Deep learning	GPU	Deep	Initial	Batch	Epoch	Normalization
---------------	-----	------	---------	-------	-------	---------------

frameworks		learning backbone	learning rate	size		mean/standard deviation
PyTorch	GTX 1080ti; NVIDIA	Resnet-50	0.01	20*60	2000	55/40

**Supplementary Table 4. Formulae of the evaluation metrics.**

Evaluation metrics	Formula
Accuracy	$Accuracy = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{TN} + N_{FN}}$
Sensitivity	$Sensitivity = \frac{N_{TP}}{N_{TP} + N_{FN}}$
Specificity	$Specificity = \frac{N_{TN}}{N_{TN} + N_{FP}}$
Positive predictive value	$Positive\ predictive\ value = \frac{N_{TP}}{N_{TP} + N_{FP}}$
Negative predictive value	$Negative\ predictive\ value = \frac{N_{TN}}{N_{TN} + N_{FN}}$

$N_{TP}$ = true positive number;  $N_{TN}$ = true negative number;  $N_{FN}$ = false positive number;  $N_{FP}$ = false negative number.

**Supplementary Table 5. Detailed baseline characteristics of the participants with histologically confirmed GISTs or GILs.**

	<b>Training set(n=493)</b>	<b>Validation set (n=105)</b>	<b>Internal test set (n=104)</b>	<b>External test set (n=50)</b>	<b>Retrospective diagnostic test set (n = 84)</b>	<b>Prospective diagnostic test set* (n = 77)</b>
Number (GIST/GIL)	227/266	47/58	50/54	24/26	30/54	36/41
Age (year)	53.9±10.7	54.3±11.2	54.1±11.6	58.6±8.7	55.1±12	57.7±8.5
Sex (male/female)	220/273	48/57	47/57	22/28	42/42	34/43
<b>Tumor diameter (mm; GIST/GIL)</b>						
<10	35/128	7/31	11/28	7/16	7/27	4/19
≥10, <20	87/91	17/17	17/17	9/5	15/21	15/12
≥20, <30	49/21	9/4	9/3	5/3	2/5	8/6
≥30, <40	27/8	7/3	5/2	2/1	4/0	6/3
≥40, <50	14/10	3/1	5/2	1/1	2/0	0/1
≥50	15/8	4/2	3/2	0/0	0/1	3/0
<b>Tumor location (GIST/GIL)</b>						
Esophagus	3/139	0/34	0/30	0/12	0/29	0/16
Stomach	214/122	45/23	47/22	24/11	30/24	32/25
Gastric cardia	8/30	2/9	3/5	1/0	1/4	1/6
Gastric fundus	102/33	20/6	20/7	7/6	11/4	10/4
Gastric body	73/40	20/8	17/10	14/5	13/13	15/14
Gastric horn	6/0	0/0	2/0	0/0	0/0	2/0
Gastric antrum	15/6	2/0	4/0	0/0	4/1	2/0

Fundus-body junction	6/9	0/0	0/0	1/0	0/1	1/0
Body-antrum junction	2/1	1/0	0/0	1/0	0/0	0/0
Multiple locations	2/3	0/0	1/0	0/0	1/1	1/1
Duodenum	7/2	1/0	0/1	0/2	0/0	2/0
Colon	0/2	0/1	1/0	0/1	0/1	0/0
Rectum	3/1	1/0	2/1	0/0	0/0	2/0

\* The prospective participants who were diagnosed as GISTs or GILs through both EUS examination and histological examination.

**Supplementary Table 6. Details of the performance of the AI model in the internal and external test sets and results of independent diagnosis and joint diagnosis in distinguishing GISTs and GILs.**

	Inner test set (AI system)		External test set (AI system)		Retrospective test set (Endosono- graphers)		Retrospective test set (AI system)		Prospective test set* (independent diagnosis)		Prospective test set* (joint diagnosis)	
	GIST	GIL	GIST	GIL	GIST	GIL	GIST	GIL	GIST	GIL	GIST	GIL
<b>Number (Ture/total)</b>	49/50	51/54	11/24	22/26	18/30	41/54	27/30	52/54	32/36	27/41	32/36	39/41
<b>Tumor diameter (mm)</b>												
<10	10/11	26/28	4/7	14/16	3/7	21/27	7/7	26/27	4/4	12/19	3/4	18/19
≥10, <20	17/17	16/17	4/9	4/5	8/15	15/21	13/15	20/21	12/15	7/12	13/15	11/12
≥20, <30	9/9	3/3	2/5	2/3	2/2	4/5	2/2	5/5	7/8	4/6	7/8	6/6
≥30, <40	5/5	2/2	1/2	1/1	4/4	0/0	3/4	0/0	6/6	3/3	6/6	3/3
≥40, <50	5/5	2/2	0/1	1/1	1/2	0/0	2/2	0/0	0/0	1/1	0/0	1/1
≥50	3/3	2/2	0/0	0/0	0/0	1/1	0/0	1/1	3/3	0/0	3/3	0/0
<b>Tumor location</b>												
Esophagus	0/0	29/30	0/0	12/12	0/0	29/29	0/0	29/29	0/0	16/16	0/0	16/16
Stomach	46/47	20/22	11/24	7/11	18/30	11/24	27/30	22/24	28/32	11/25	29/32	23/25
Gastric cardia	3/3	5/5	0/1	0/0	1/1	3/4	0/1	3/4	1/1	5/6	0/1	6/6
Gastric fundus	20/20	7/7	3/7	4/6	7/11	3/4	11/11	4/4	8/10	1/4	9/10	4/4
Gastric body	16/17	8/10	7/14	3/5	7/13	4/13	12/13	12/13	13/15	5/14	14/15	12/14

Gastric horn	2/2	0/0	0/0	0/0	0/0	0/0	0/0	0/0	2/2	0/0	2/2	0/0
Gastric antrum	4/4	0/0	0/0	0/0	2/4	0/1	3/4	1/1	2/2	0/0	2/2	0/0
Fundus-body junction	0/0	0/0	1/1	0/0	0/0	0/1	0/0	1/1	1/1	0/0	1/1	0/0
Body-antrum junction	0/0	0/0	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Multiple locations	1/1	0/0	0/0	0/0	1/1	1/1	1/1	1/1	1/1	0/1	1/1	1/1
Duodenum	0/0	1/1	0/0	2/2	0/0	0/0	0/0	0/0	2/2	0/0	1/2	0/0
Colon	1/1	0/0	0/0	1/1	0/0	1/1	0/0	1/1	0/0	0/0	0/0	0/0
Rectum	2/2	1/1	0/0	0/0	0/0	0/0	0/0	0/0	2/2	0/0	2/2	0/0

\* The prospective participants who were diagnosed as GISTs or GILs through both EUS examination and histological examination.

**Supplementary Table 7. Pixel values analysis in different centers.**

Hospital (number of tumor images)	Pixel value mean of tumor images	Pixel value standard deviation of tumor images
AHQSN (4000)	50.7	31.3
GIST (2000)	49.5	29.2
GIL (2000)	52.0	33.6
AHQSL (210)	51.6	30.2
GIST (105)	49.6	30.6
GIL (105)	53.7	29.9
QLH (154)	61.0	35.4
GIST (77)	76.9	36.6
GIL (77)	44.9	33.8
QMEH (66)	17.3	19.4
GIST (33)	8.5	13.3
GIL (33)	26.0	25.5

AHQSN = Affiliated Hospital of Qingdao University Shinan Hospital,

AHQSL = Affiliated Hospital of Qingdao University Laoshan Hospital,

QLH = Qilu Hospital,

QMEH = Qingdao Municipal Hospital (Group) East Hospital.



**Supplementary Table 8: Consecutive subjects in the process of collecting the diagnostic tests.**

	Retrospective diagnostic test			Prospective diagnostic test		
<b>Endoscopic center</b>	AHQSN	AHQLS	AHQHD	AHQSN	AHQLS	AHQHD
<b>Consecutive subjects</b>	254	131	147	224	158	126
<b>Participants with clear histological diagnosis</b>	54	33	39	45	56	31
<b>GISTs</b>	13	7	10	16	12	8
<b>GILs</b>	20	15	19	12	22	10
<b>Other SELs</b>	21	11	10	17	22	13
<b>Participants without histological examination</b>	200	97	107	179	98	94
<b>Participants with unclear histological diagnosis</b>	0	1	1	0	4	1

**Supplementary Table 9. Details of the cases diagnosed with GISTs or GILs on EUS by endosonographers in AHQSN in the process of collecting the internal dataset.**

Cases diagnosed with GISTs or GILs on EUS by endosonographers (N=2264)			
Cases with histological reports			Cases without histological reports
Number	699		1565
Diameter (mm; <20/≥20/no records)	466/232/1		1444/119/2
Histologically confirmed GISTs or GILs		Other SELs or SELs without clear histological diagnosis	
Number	656	43	
Histologically confirmed GIST/GIL	310/346	-	
GIST/GIL diagnosed on EUS	365/291	36/7	
Consistency rate	76.2%	0%	
Diameter (<20mm/≥20mm/no records)	452/203/1	14/29/0	
		Inclusion	Exclusion
Number		648	8
Histologically confirmed GIST/GIL		307/341	3/5
GIST/GIL diagnosed on EUS		362/286	3/5
Consistency rate		75.9%	100%
Diameter (<20mm/≥20mm/no records)		446/202/0	6/1/1
		Histologically confirmed GISTs	Histologically confirmed GILs
Number		307	341
GIST/GIL diagnosed on EUS		257/50	105/236

<b>Consistency rate</b>	83.7%	68.9%			
<b>Diameter (&lt;20mm/≥20mm)</b>	167/140	279/62			

**Supplementary Table 10. Histological diagnosis of EUS subjects with histologically-confirmed SELs (not GISTs or GILs) or the EUS subjects without clear histological diagnosis in AHQSN in the process of collecting of the internal dataset.**

<b>Histologically diagnosis</b>	<b>N = 38</b>
<b>Schwannoma</b>	16
<b>Cyst</b>	3
<b>Metastasis</b>	3
<b>Neurofibroma</b>	3
<b>Glomus tumor</b>	2
<b>Neuroendocrine tumor</b>	2
<b>Calcifying fibrous tumor</b>	1
<b>Ectopic pancreas</b>	1
<b>Granular cell tumor</b>	1
<b>Inflammatory fibroid polyp</b>	1
<b>Myoepithelial hamartoma</b>	1
<b>Squamous cell carcinoma</b>	1
<b>Spindle cell tumor (no clear histological diagnosis)</b>	3

**Supplementary Table 11. Histological diagnosis of the subjects with histologically-confirmed GISTs or GILs who were wrongly diagnosed as other kinds of SELs in AHQSN in the process of collecting internal dataset.**

	N =54
<b>Histologically confirmed GIST/GIL</b>	17/37
<b>Diameter (&lt;20mm/≥20mm/no records)</b>	40/14/0
<b>EUS diagnosis</b>	
Granular cell tumor	16
Neuroendocrine tumor	12
Inflammatory fibroid polyp	4
Fibroma	4
Ectopic pancreas	3
Polyp	3
Cyst	2
Hamartoma	2
Gastric cancer/	1
Glomus tumor	1
Schwannoma	1
Hemangioma	1
Unclear diagnosis or no diagnosis	4

**Supplementary Table 12. Running time of the AI system on workstations with different CPUs.**

CPU	AMD Ryzen 7-2700	Intel Core i3-8100	Intel Core i5-4590
Running time (second/5 images)	<3	<6	<10

## Reference

1. HAJIAN-TILAKI K. Sample size estimation in diagnostic test studies of biomedical informatics. *J Biomed Inform* 2014;48:193-204.
2. O'SULLIVAN JW, BANERJEE A, HENEGHAN C, et al. Verification bias. *BMJ Evid Based Med* 2018;23:54-55.
3. SCHMIDT RL, JEDRZKIEWICZ JD, ALLRED RJ, et al. Verification bias in diagnostic accuracy studies for fine- and core needle biopsy of salivary gland lesions in otolaryngology journals: a systematic review and analysis. *Head Neck* 2014;36:1654-61.
4. KOSINSKI AS, BARNHART HX. A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. *Stat Med* 2003;22:2711-21.
5. KOSINSKI AS, BARNHART HX. Accounting for nonignorable verification

- bias in assessment of diagnostic tests. *Biometrics* 2003;59:163-71.
6. ZHOU X-H. Comparing accuracies of two screening tests in a two-phase study for dementia. 1998;47:135-147.
  7. DENDUKURI N, JOSEPH L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 2001;57:158-67.
  8. BAKER SG. Evaluating multiple diagnostic tests with partial verification. *Biometrics* 1995;51:330-7.
  9. STANDARDS OF PRACTICE C, FAULX AL, KOTHARI S, et al. The role of endoscopy in subepithelial lesions of the GI tract. *Gastrointest Endosc* 2017;85:1117-1132.

**Table 1. Baseline characteristics of the internal and external datasets.**

	<b>Training set (n=493)</b>	<b>Validation set (n=105)</b>	<b>Internal test set (n=104)</b>	<b>External test set (n=50)</b>
<b>Number (GIST/GIL)</b>	227/266	47/58	50/54	24/26
<b>Age (year)</b>	53.9±10.7	54.3±11.2	54.1±11.6	58.6±8.7
<b>Sex (male/female)</b>	220/273	48/57	47/57	22/28
<b>Tumor diameter (mm; GIST/GIL)</b>				
<20	122/219	24/48	28/45	16/21
≥20	105/47	23/10	22/9	8/5
<b>Tumor location (GIST/GIL)</b>				
<b>Esophagus</b>	3/139	0/34	0/30	0/12
<b>Stomach</b>	214/122	45/23	47/22	24/11
<b>Duodenum</b>	7/2	1/0	0/1	0/2
<b>Colon</b>	0/2	0/1	1/0	0/1
<b>Rectum</b>	3/1	1/0	2/1	0/0

GIST=gastrointestinal stromal tumor, GIL=gastrointestinal leiomyoma.

**Table 2. Baseline characteristics of participants with histologically-confirmed GIST or GIL in the diagnostic tests.**

	Retrospective diagnostic test			Prospective diagnostic test*		
Endoscopic center	AHQSN	AHQLS	AHQHD	AHQSN	AHQLS	AHQHD
Number	13/20	7/15	10/19	16/10	12/22	8/9
(GIST/GIL)						
Sex (male/female)	16/17	10/12	16/13	13/13	12/22	9/8
Age(year)	56.1 ± 12.7	54.9 ± 9.9	54.1 ± 13	58 ± 8.5	58.1 ± 8.5	56.2 ± 9
Lesion diameter (mm; GIST/GIL)						
<20	12/18	3/14	7/16	10/8	5/19	4/4
≥20	1/2	4/1	3/3	6/2	7/3	4/5
Lesion location (GIST/GIL)						
Esophagus	0/10	0/10	0/9	0/5	0/7	0/4
Stomach	13/10	7/5	10/9	13/5	12/15	7/5
Duodenum	0/0	0/0	0/0	1/0	0/0	1/0
Colon	0/0	0/0	0/1	0/0	0/0	0/0
Rectum	0/0	0/0	0/0	2/0	0/0	0/0

\* The prospective participants who were simultaneously diagnosed as GISTs or GILs through EUS examination and histological examination.



**Table 3. Histological diagnoses of the prospective participants.**

Diagnosis on EUS	Independent diagnosis by endosonographers				Diagnosis with the AI system	
	GIST	GIL	Other kinds of SELs	No clear diagnosis	GIST	GIL
Number	127	129	241	11	77	179
<b>Histological diagnosis</b>	50	31	54	2	36	45
GIST	32	4	0	0	32	4
GIL	14	27	3	0	2	39
Other kinds of SELs	4	0	46	2	2	2
No clear diagnosis	0	0	5	0	0	0



