Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study

Huiyan Luo, Guoliang Xu*, Chaofeng Li*, Longjun He*, Linna Luo*, Zixian Wang*, Bingzhong Jing, Yishu Deng, Ying Jin, Yin Li, Bin Li, Wencheng Tan, Caisheng He, Sharvesh Raj Seeruttun, Qiubao Wu, Jun Huang, De-wang Huang, Bin Chen, Shao-bin Lin, Qin-ming Chen, Chu-ming Yuan, Hai-xin Chen, Heng-ying Pu, Feng Zhou, Yun He, Rui-hua Xu*

Summary

Background Upper gastrointestinal cancers (including oesophageal cancer and gastric cancer) are the most common cancers worldwide. Artificial intelligence platforms using deep learning algorithms have made remarkable progress in medical imaging but their application in upper gastrointestinal cancers has been limited. We aimed to develop and validate the Gastrointestinal Artificial Intelligence Diagnostic System (GRAIDS) for the diagnosis of upper gastrointestinal cancers through analysis of imaging data from clinical endoscopies.

Methods This multicentre, case-control, diagnostic study was done in six hospitals of different tiers (ie, municipal, provincial, and national) in China. The images of consecutive participants, aged 18 years or older, who had not had a previous endoscopy were retrieved from all participating hospitals. All patients with upper gastrointestinal cancer lesions (including oesophageal cancer and gastric cancer) that were histologically proven malignancies were eligible for this study. Only images with standard white light were deemed eligible. The images from Sun Yat-sen University Cancer Center were randomly assigned (8:1:1) to the training and intrinsic verification datasets for developing GRAIDS, and the internal validation dataset for evaluating the performance of GRAIDS. Its diagnostic performance was evaluated using an internal and prospective validation set from Sun Yat-sen University Cancer Center (a national hospital) and additional external validation sets from five primary care hospitals. The performance of GRAIDS was also compared with endoscopists with three degrees of expertise: expert, competent, and trainee. The diagnostic accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of GRAIDS and endoscopists for the identification of cancerous lesions were evaluated by calculating the 95% CIs using the Clopper-Pearson method.

Findings 1 036 496 endoscopy images from 84 424 individuals were used to develop and test GRAIDS. The diagnostic accuracy in identifying upper gastrointestinal cancers was 0·955 (95% CI 0·952–0·957) in the internal validation set, 0·927 (0·925–0·929) in the prospective set, and ranged from 0·915 (0·913–0·917) to 0·977 (0·977–0·978) in the five external validation sets. GRAIDS achieved diagnostic sensitivity similar to that of the expert endoscopist (0·942 [95% CI 0·924–0·957] *vs* 0·945 [0·927–0·959]; p=0·692) and superior sensitivity compared with competent (0·858 [0·832–0·880], p<0·0001) and trainee (0·722 [0·691–0·752], p<0·0001) endoscopists. The positive predictive value was 0·814 (95% CI 0·788–0·838) for GRAIDS, 0·932 (0·913–0·948) for the expert endoscopist, 0·974 (0·960–0·984) for the competent endoscopist, and 0·824 (0·795–0·850) for the trainee endoscopist. The negative predictive value was 0·978 (95% CI 0·971–0·984) for GRAIDS, 0·980 (0·974–0·985) for the expert endoscopist, 0·951 (0·942–0·959) for the competent endoscopist, and 0·904 (0·893–0·916) for the trainee endoscopist.

Interpretation GRAIDS achieved high diagnostic accuracy in detecting upper gastrointestinal cancers, with sensitivity similar to that of expert endoscopists and was superior to that of non-expert endoscopists. This system could assist community-based hospitals in improving their effectiveness in upper gastrointestinal cancer diagnoses.

Funding The National Key R&D Program of China, the Natural Science Foundation of Guangdong Province, the Science and Technology Program of Guangdong, the Science and Technology Program of Guangzhou, and the Fundamental Research Funds for the Central Universities.

Lancet Oncol 2019; 20: 1645–54

Published **Online**
October 4, 2019
https://doi.org/10.1016/S1470-2045(19)30637-0

See **Comment** page 1616

*Contributed equally

Department of Medical Oncology (Prof H Luo MD, L Luo MD, Z Wang MD, Y Jin MD, Prof R-h Xu MD), **Department of Endoscopy** (Prof G Xu MD, L He MD, Y Li MD, W Tan MD), **Artificial Intelligence Laboratory** (C Li PhD, B Jing MS, Y Deng MS, B Li MS, C He PhD), **Department of Gastric Surgery** (S R Seeruttun PhD), **and Medical Administration Department** (H-y Pu PhD, F Zhou PhD, Y He PhD), **State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China; Department of Endoscopy, Jiangxi Cancer Hospital, Nanchang, China** (Prof Q Wu MD, J Huang MD); **Department of Digestive Internal, Wuzhou Red Cross Hospital, Wuzhou, China** (Prof D-w Huang MD); **Department of Digestive Internal, The North Guangdong People's Hospital, Shaoguan, China** (Prof B Chen MD); **Department of Digestive Internal, Puning People's Hospital, Puning, China** (Prof S-b Lin MD, Prof Q-m Chen MD); **and Department of Digestive Internal, Jieyang People's Hospital, Jieyang, China** (Prof C-m Yuan MD, Prof H-x Chen MD)

Introduction

Upper gastrointestinal cancers (including oesophageal cancer and gastric cancers) are among the most common malignancies and causes of cancer-related deaths worldwide,[1] representing a great challenge to health-care systems because of their aggressive presentation.[2] Most upper gastrointestinal cancers are diagnosed at advanced stages because their signs and symptoms tend to be latent and non-specific, leading to an overall poor prognosis, but if detected early, 5-year survival can exceed 90%.[3–5]

Correspondence to:
Dr Rui-hua Xu, Department of Medical Oncology, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
xurh@sysucc.org.cn

Research in context

Evidence before this study

We searched PubMed for publications on artificial intelligence-based endoscopy diagnostic systems for upper gastrointestinal cancers published from database inception to Dec 31, 2018, using search terms "artificial intelligent" or "AI", "endoscopy", and "upper gastrointestinal cancer" or "esophageal cancer" or "gastric cancer", without language restrictions. The results of the scientific literature search, however, were limited. Despite some encouraging preliminary reports, definitive conclusions from the existing publications on the clinical applicability and reliability of an artificial intelligence-assisted endoscopic diagnosis remained questionable due to the retrospective nature, single-disease investigations, low sample size, and either inadequate or no validation of these studies. Until now, the clinical use of a real-time artificial intelligence image recognition system for upper gastrointestinal cancer detection remained investigational.

Added value of this study

We developed and validated the Gastrointestinal Artificial Intelligence Diagnosis System (GRAIDS), a deep learning semantic segmentation model capable of providing real-time automated detection of upper gastrointestinal cancers, from suspicious lesions during endoscopic examinations based on 1 036 496 endoscopy images from 84 424 individuals from different tier hospitals across China. The diagnostic performance of GRAIDS was also evaluated in less experienced hospitals with low volumes of patients with upper gastrointestinal cancers. GRAIDS was able to detect upper gastrointestinal cancer, at a latency of less than 40 ms in real-time imaging analysis, with high diagnostic accuracy across all participating hospitals. When compared with endoscopists who had different degrees of expertise, GRAIDS demonstrated sensitivity comparable to that of expert endoscopists and was superior to that of nonexperts. To promote the clinical applicability of GRAIDS, a cloud-based multi-institutional artificial intelligence platform has been designed to provide real-time assistance for diagnosing upper gastrointestinal cancers during endoscopic examinations and for retrospectively assessing endoscopic images, acting as a second expert opinion to diagnose or decrease the risk of missing suspicious lesions.

Implications of all the available evidence

To our knowledge, GRAIDS is the first real-time artificial intelligence-aided image recognition system that has been implemented in clinical practice for detecting upper gastrointestinal cancers during endoscopy. It maintained a robust diagnostic performance even after validation in different tier hospitals. GRAIDS can further assist non-expert endoscopists from primary basic or low-volume hospitals to improve their diagnostic accuracy of upper gastrointestinal cancers, similar to the level of expert endoscopists, and thereby providing both real-time and retrospective opportunities for all hospitals to improve the effectiveness of diagnosis and screening for upper gastrointestinal cancers.

To overcome this challenge, upper gastrointestinal endoscopic strategies and techniques, such as narrow-band imaging and confocal laser endomicroscopy, have been developed and implemented in many countries, resulting in an increased detection of early upper gastrointestinal cancer and a decrease in mortality.[6–10] However, the risk of missing suspicious upper gastrointestinal cancers in endoscopy examinations might still be high in hospitals with low patient volume, in less developed or remote regions, and even in countries where many endoscopies are performed.[11]

Artificial intelligence has already shown potential for assisting humans in various medical fields.[12–14] For example, artificial intelligence can automatically extrapolate complex microimaging structures (ie, the extent of mucosal-tubular branches and colour intensity anomalies) and identify quantitative pixel-level features,[15] which are often undetectable by the human eye. Clinically, the most important use of endoscopic artificial intelligence is to assist in differentiation between neoplastic and non-neoplastic lesions. Although encouraging preliminary results have been published regarding the use of artificial intelligence in the diagnosis of upper gastrointestinal cancers,[16–18] their clinical impact has mostly been minimal because of study design (ie, single-centre studies, small sample sizes, and post-hoc analyses).

We aimed to develop a diagnostic platform to detect upper gastrointestinal cancers using real-world endoscopic imaging data from six hospitals with varying experience in the endoscopic diagnosis of upper gastrointestinal cancer.

Methods

Study design and participants

This multicentre, case-control, diagnostic study was done in six hospitals in China. We retrospectively obtained endoscopic images for the development and validation of the Gastrointestinal Artificial Intelligence Diagnostic System (GRAIDS) from the imaging database at Sun Yat-sen University Cancer Center (SYSUCC; Guangzhou, China), a national hospital.

From July 21, 2018, GRAIDS was published online and implemented in SYSUCC's endoscopic practice. A monitor for real-time analysis using GRAIDS was fixed adjacent to the original endoscopy monitor and independent cohorts of consecutive participants receiving upper gastrointestinal endoscopy were prospectively enrolled. These participants were defined as the prospective validation set.

To generalise the applicability of GRAIDS in clinical practice, endoscopic images were also obtained from five municipal or provincial hospitals across China:

North Guangdong People's Hospital, Shaoguan; Wuzhou's Red Cross Hospital, Wuzhou; Jiangxi Cancer Hospital, Nanchang; Puning People's Hospital, Puning; and Jieyang People's Hospital, Jieyang.

All endoscopies were carried out for screening or pretreatment examination in daily clinical practice. The images from consecutive participants, aged 18 years or older, who had not had an endoscopy previously were retrieved from all participating hospitals. All patients with upper gastrointestinal cancer lesions (including oesophageal cancer and gastric cancer) that were histologically proven malignancies were eligible for this study. We excluded participants with a history of cancer or gastrointestinal surgery, and those with a biopsy for upper gastrointestinal lesions without a definitive pathological diagnosis or pathological report. Board-certified pathologists did the pathological assessments based on the WHO Classification of Tumours at individual sites. All pathological assessments were based on haematoxylin-stained and eosin-stained tissue slides. For participants without cancer (normal controls or those with histologically confirmed benign tumours), there were no specific exclusion criteria regarding demographic or clinical features.

This study was approved by the relevant independent institutional review boards of each participating hospital and performed according to the Helsinki declaration. All patients from the SYSUCC prospective validation set provided written informed consent before participation. For patients whose endoscopic images were stored in the retrospective databases at each participating hospital, informed consent was exempted by the institutional review boards of the participating hospitals.

Endoscopy and image quality control

All images were captured in high-resolution but with different endoscopes (GIF-HQ290, GIF-H260, GIF-Q240, GIF-Q180, GIF-H170, or GIF-LV1, Olympus Medical Systems, Tokyo, Japan; or EG590WR, EG600ZW, EG-L590ZW, or EG-760Z, Fujifilm Medical Systems, Shanghai, China) and video systems (EVIS LUCERA CV-260/CLV-260, EVIS LUCERA ELITE CV-290/CLV-290SL; Olympus Medical Systems). All upper gastrointestinal endoscopic images were stored in a jpeg format in the imaging databases at the six hospitals. Only images with standard white light were deemed eligible. Dye stained images, narrow-band imaging, and poor-quality images resulting from halation, blurs, defocus, mucus, and poor air insufflation, as well as non-endoscopic images, were excluded.

Eight highly experienced endoscopists from SYSUCC, each of whom had a minimum of 5 years of experience and had performed more than 3000 examinations, assessed the quality of all the images. All of the upper gastrointestinal cancer lesions were labelled manually by the same group of endoscopists. They carefully marked the border of each cancer lesion. Those endoscopic images that did not match with the pathological reports in terms of anatomical locations were discarded. An equal number of images from the six hospitals were assigned to four groups of experienced endoscopists (two endoscopists in each group) for quality control, labelling, and delineation. The two endoscopists in the same group cooperated in labelling and delineation. Regarding delineation, one endoscopist implemented the delineation under the supervision of the other. The image selection, labelling, and delineation were finalised only when the two endoscopists from the same group reached a consensus.

Development of the GRAIDS algorithm

The images from SYSUCC were randomly assigned (8:1:1) to the training and intrinsic verification datasets for developing GRAIDS, and the internal validation dataset for evaluating the performance of GRAIDS. The GRAIDS' algorithm was based on the concept of DeepLab's V3+ (released by Google [Mountain View, CA, USA] in 2018),[19,20] and comprised an encoder and decoder module (appendix pp 3–5). The model had one input and two outputs. The input of the model was the endoscopic images of the upper gastrointestinal tract. The first output was a standard two-class task for determining whether the input picture contained a tumour. The second output implemented a segmentation task that captured the tumour region of the input image. The labelling and delineation data from the four endoscopist groups (each comprising two endoscopists) were adopted as the gold standard in training GRAIDS. A learning curve measured the image classification and an intersection-over-union (IOU) measured the image segmentation performance of the model (appendix pp 3–4, 6). See the appendix (pp 3–4) for additional details on the deep learning algorithm.

See Online for appendix

Validation of the GRAIDS algorithm

We first validated the performance of GRAIDS in the identification of upper gastrointestinal cancers in patients using an internal validation dataset and a prospective validation dataset from SYSUCC. We then assessed the robustness of GRAIDS using external validation datasets from the five participating hospitals, each with a low volume of patients with upper gastrointestinal cancers.

For further performance evaluation, we randomly selected a subset of patients' images with histologically confirmed upper gastrointestinal cancers from the prospective validation set. Three endoscopists, of varying degrees of expertise (expert, competent, and trainee), who were masked to the patients' demographics and final histopathological results, were asked to independently complete the same test images, and their results were compared with those of GRAIDS. All three endoscopists were not involved in the selection

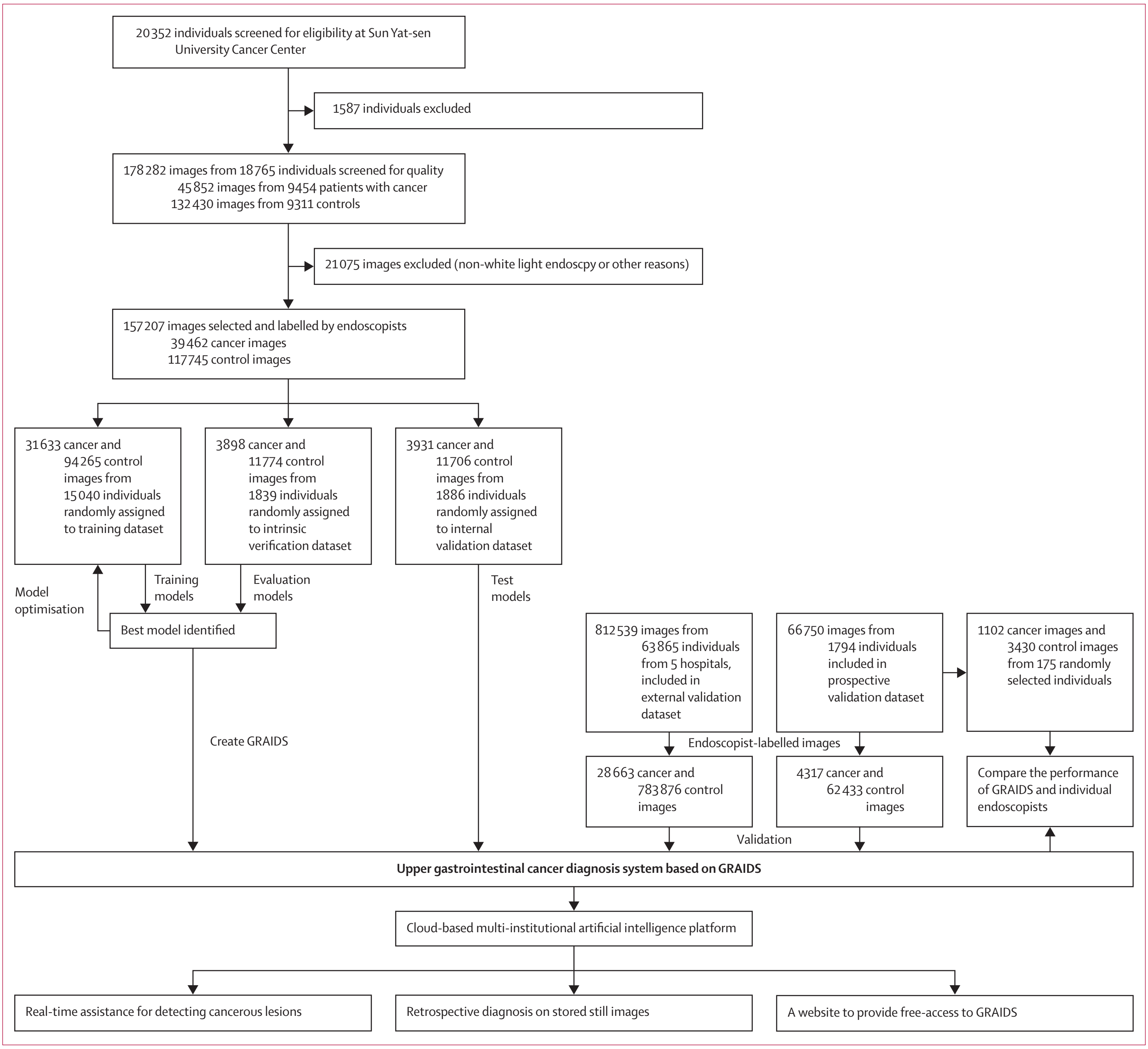
20 352 individuals screened for eligibility at Sun Yat-sen University Cancer Center
1587 individuals excluded
178 282 images from 18 765 individuals screened for quality
45 852 images from 9454 patients with cancer
132 430 images from 9311 controls
21 075 images excluded (non-white light endoscpy or other reasons)
157 207 images selected and labelled by endoscopists
39 462 cancer images
117 745 control images
31 633 cancer and 94 265 control images from 15 040 individuals randomly assigned to training dataset
3898 cancer and 11 774 control images from 1839 individuals randomly assigned to intrinsic verification dataset
3931 cancer and 11 706 control images from 1886 individuals randomly assigned to internal validation dataset
Model optimisation
Training models
Evaluation models
Test models
Best model identified
Create GRAIDS
812 539 images from 63 865 individuals from 5 hospitals, included in external validation dataset
66 750 images from 1794 individuals included in prospective validation dataset
1102 cancer images and 3430 control images from 175 randomly selected individuals
Endoscopist-labelled images
28 663 cancer and 783 876 control images
4317 cancer and 62 433 control images
Compare the performance of GRAIDS and individual endoscopists
Validation
Upper gastrointestinal cancer diagnosis system based on GRAIDS
Cloud-based multi-institutional artificial intelligence platform
Real-time assistance for detecting cancerous lesions
Retrospective diagnosis on stored still images
A website to provide free-access to GRAIDS

***Figure 1:* Workflow diagram for the development and evaluation of GRAIDS**

Control images were images with benign lesions or images that did not show any evidence associated with upper gastrointestinal malignancy in clinical testing, as determined by the endoscopists. GRAIDS=Gastrointestinal Artificial Intelligence Diagnosis System.

and labelling of the images, and the images were also mixed up and de-identified before the endoscopists' assessments. The expert endoscopist was a professor with more than 10 years of experience in endoscopic procedures. The competent endoscopist was an attending doctor with more than 5 years of experience who had finished both clinical and specific endoscopic training. The trainee was a resident with 2 years of endoscopic experience.

Statistical analysis

The diagnostic accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of GRAIDS for the identification of cancerous lesions were evaluated (appendix p 7) by calculating the 95% CIs using the Clopper-Pearson method. We used the receiver operating characteristic (ROC) curve to show the diagnostic ability of the deep learning algorithm in discriminating patients with upper gastrointestinal cancer from controls. ROC curves were created by plotting the proportion of true positive cases (sensitivity) against the proportion of false positive cases (1−specificity), by varying the predictive probability threshold. A larger the area under the ROC curve (AUC) indicated better diagnostic performance. All statistical tests were two-sided with a significance level of 0·05. Statistical analyses were done with R software, version 3.5.1.

Role of the funding source

The funder had no role in this study's design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in this study and had final responsibility over the decision to submit it for publication.

Results

Between Jan 12, 2009, and Sept 30, 2017, 314726 images from 20 352 participants were obtained from the upper gastrointestinal endoscopic imaging database at SYSUCC (figure 1). Owing to undetermined pathological diagnosis and unavailable pathological reports, 1587 (7·8%) of 20 352 participants were excluded. After quality control evaluation, 21075 (11·8%) of 178 282 images were discarded because they were non-endoscopic images of poor quality or images inconsistent with pathological reports in terms of anatomical locations. For patients with cancer, only images of cancer lesions were included (n=39 462). For participants without cancer, 117745 images were used as the control group (figure 1). For the prospective validation dataset, 4317 cancer images and 62 433 control images were prospectively collected and labelled at SYSUCC between July 21, 2018, and Nov 20, 2018.

At the five other participating hospitals, between July 21, 2018, and Nov 20, 2018, 2439 cancer and 73 015 control images were obtained from North Guangdong People's Hospital, 5244 cancer and 197 588 control images from Wuzhou Red Cross Hospital, 9712 cancer and 112 185 control images from Jiangxi Cancer Hospital, 7095 cancer and 286 095 control images from Puning People's Hospital, and 4173 cancer and 114 993 control images from Jieyang People's Hospital. Overall, 1 036 496 endoscopy images from 84 424 individuals were used to develop and test GRAIDS.

The prevalence of upper gastrointestinal cancer was 50·2% (7557 of 15 040 patients) in the training set, 51·0% (938 of 1839 patients) in the intrinsic verification set, 50·8% (959 of 1886 patients) in the internal validation set, 32·0% (574 of 1794 patients) in the prospective validation set, 9·2% (794 of 8634 patients) in the Jiangxi Cancer Hospital external validation set, 9·5% (390 of 4109) in North Guangdong People's Hospital, 4·8% (830 of 17 293 patients) in Wuzhou Red Cross Hospital, 3·8% (993 of 26 143 patients) in Puning People's Hospital, and 7·2% (552 of 7686 patients) in Jieyang People's Hospital (table 1). Detailed staging information for

	Sun Yat-sen University Cancer Center validation (n=20 559)				External validation (n=63 865)					p value*
	Training (n=15 040)	Verification (n=1839)	Internal (n=1886)	Prospective (n=1794)	JCH (n=8634)	NGPH (n=4109)	WCH (n=17 293)	PPH (n=26 143)	JPH (n=7686)	
Age, years	55·55 (12·43)	56·08 (12·71)	55·27 (12·16)	52·91 (11·84)	51·64 (13·82)	49·44 (14·56)	47·27 (14·91)	42·77 (15·57)	46·10 (15·79)	<0·0001
Sex	..	..	..	..	..	..	..	..	..	<0·0001
Male	9450 (62·8%)	1181 (64·2%)	1173 (62·2%)	1034 (57·6%)	4344 (50·3%)	1849 (45·0%)	7793 (45·1%)	11 241 (43·0%)	3907 (50·8%)	..
Female	5590 (37·2%)	658 (35·8%)	713 (37·8%)	760 (42·4%)	4290 (49·7%)	2260 (55·0%)	9500 (54·9%)	14 902 (57·0%)	3779 (49·2%)	..
Gastric cancer	2729 (18·1%)	347 (18·9%)	338 (17·9%)	172 (9·6%)	457 (5·3%)	244 (5·9%)	329 (1·9%)	264 (1·0%)	206 (2·7%)	..
Oesophageal cancer	4091 (27·2%)	504 (27·4%)	537 (28·5%)	333 (18·6%)	301 (3·5%)	137 (3·3%)	470 (2·7%)	719 (2·8%)	322 (4·2%)	..
Other cancer†	737 (4·9%)	87 (4·7%)	84 (4·5%)	69 (3·8%)	36 (0·4%)	9 (0·2%)	31 (0·2%)	10 (0·0%)	24 (0·3%)	..
Benign disease	4443 (29·5%)	513 (27·9%)	542 (28·7%)	774 (43·1%)	3441 (39·9%)	2232 (54·3%)	8989 (52·0%)	12 662 (48·4%)	2378 (30·9%)	..
No disease	3040 (20·2%)	388 (21·1%)	385 (20·4%)	446 (24·9%)	4399 (50·9%)	1487 (36·2%)	7474 (43·2%)	12 488 (47·8%)	4756 (61·9%)	..

Data are mean (SD) or n (%). JCH=Jiangxi Cancer Hospital. NGPH=North Guangdong People's Hospital. WCH=Wuzhou Red Cross Hospital. PPH=Puning People's Hospital. JPH=Jieyang People's Hospital. *p<0·05 indicates that patient age and sex composition or the prevalence of malignancy varied significantly by hospital (the Kruskal-Wallis H test was used to test whether patient age varied significantly by hospital, and the χ^2 test was used to test whether sex composition or the prevalence of malignancy varied significantly by hospital). †Other cancer includes lymphoma, sarcoma, and neuroendocrine neoplasm that occurred in the upper gastrointestinal tract.

Table 1: **Baseline characteristics**

	Sun Yat-sen University Cancer Center validation		External validation				
	Internal validation set	Prospective set	JCH	NGPH	WCH	PPH	JPH
Accuracy (95% CI)	0·955 (0·952–0·957)	0·927 (0·925–0·929)	0·915 (0·913–0·917)	0·949 (0·947–0·951)	0·977 (0·977–0·978)	0·970 (0·969–0·971)	0·947 (0·946–0·948)
Sensitivity (95% CI)	0·940 (0·932–0·947)	0·946 (0·939–0·953)	0·943 (0·938–0·948)	0·946 (0·937–0·955)	0·907 (0·899–0·915)	0·965 (0·961–0·969)	0·982 (0·978–0·986)
Specificity (95% CI)	0·961 (0·957–0·965)	0·926 (0·924–0·928)	0·913 (0·911–0·915)	0·949 (0·947–0·951)	0·979 (0·978–0·980)	0·970 (0·969–0·971)	0·945 (0·944–0·946)
Positive predictive value (95% CI)	0·889 (0·878–0·899)	0·468 (0·458–0·478)	0·484 (0·477–0·491)	0·384 (0·372–0·396)	0·538 (0·528–0·548)	0·443 (0·434–0·451)	0·394 (0·384–0·403)
Negative predictive value (95% CI)	0·979 (0·976–0·982)	0·996 (0·995–0·997)	0·995 (0·995–0·995)	0·998 (0·998–0·998)	0·997 (0·997–0·998)	0·999 (0·999–0·999)	0·999 (0·999–0·999)

GRAIDS=Gastrointestinal Artificial Intelligence Diagnosis System. JCH=Jiangxi Cancer Hospital. NGPH=North Guangdong People's Hospital. WCH=Wuzhou Red Cross Hospital. PPH=Puning People's Hospital. JPH=Jieyang People's Hospital.

***Table 2:* Performance of GRAIDS in different validation sets**

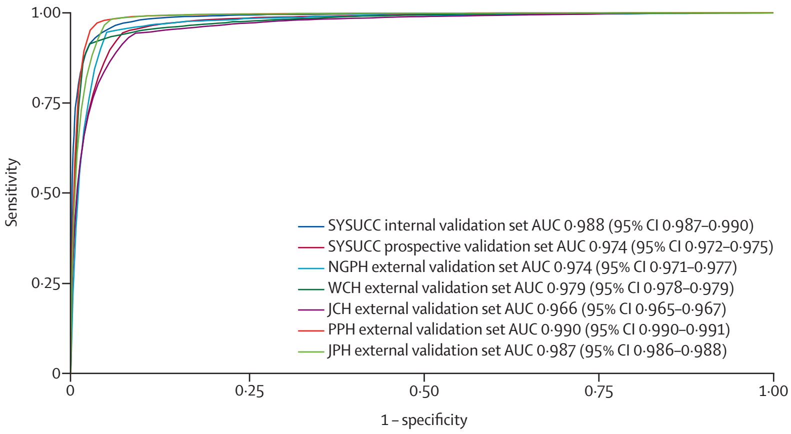

***Figure 2:* Receiver operating characteristic curves illustrate GRAIDS' ability to detect upper gastrointestinal cancer**

GRAIDS=Gastrointestinal Artificial Intelligence Diagnosis System. SYSUCC=Sun Yat-sen University Cancer Center. AUC=area under the receiver operating characteristic curve. NGPH=North Guangdong People's Hospital. WCH=Wuzhou Red Cross Hospital. JCH=Jiangxi Cancer Hospital. PPH=Puning People's Hospital. JPH=Jieyang People's Hospital.

gastrointestinal cancers was only available in the prospective validation set (appendix p 11).

After 176 epochs (iterations through the entire training set), the training procedure was concluded because of an absence of further improvement in both accuracy and cross-entropy loss in both tasks and in IOU in the second task (appendix pp 3–4, 6). There was a high degree of agreement between the predicted region by GRAIDS in segmentation of upper gastrointestinal cancer lesions and the region labelled by the endoscopists (appendix p 7). The median IOU was 0·737 (IQR 0·579–0·848) in the internal validation set.

GRAIDS was accurate in identifying patients with upper gastrointestinal cancer in all the seven validation sets (table 2). Diagnostic accuracies were 0·955 (95% CI 0·952–0·957 in the internal SYSUCC validation dataset and 0·927 (0·925–0·929) in the prospective SYSUCC validation set. Diagnostic accuracies in the external validation were 0·915 (95% CI 0·913–0·917) for the Jiangxi Cancer Hospital, 0·949 (0·947–0·951) for North Guangdong People's Hospital, 0·977 (0·977–0·978) for Wuzhou Red Cross Hospital, 0·970 (0·969–0·971) for Puning People's Hospital, and 0·947 (0·946–0·948) for Jieyang People's Hospital. Its sensitivity, specificity, and NPV were higher than 0·90 in all of the validation sets. Its PPV varied across the validation sets from 0·384 (95% CI 0·372–0·396) in the North Guangdong People's Hospital to 0·889 (0·878–0·899) in SYSUCC (table 2), but the proportion of false positive cases was less than 10% in all the validation datasets (appendix p 8). The most common cause of false positives in SYSUCC's internal validation cohort and the prospective cohort was normal anatomical structures (ie, cardia, pylorus, and angulus), as well as elevation of the gastric wall during peristalsis (appendix p 12).

Similarly, high AUC values were also observed in the five external validation datasets (ranging from 0·966 [0·965–0·967] to 0·990 [0·990–0·991]; figure 2).

The test results for GRAIDS and the endoscopists in differentiating between a subset of 4532 images (1102 [24·3%] cancer and 3430 [75·7%] control images) from the prospective validation set are shown in table 3. GRAIDS was accurate in detecting upper gastrointestinal cancer, with an accuracy of 0·928 (95% CI 0·919–0·937). Among the endoscopists, the accuracy of the expert endoscopist was significantly higher than that of GRAIDS at 0·967 (95% CI 0·961–0·973; p<0·0001), whereas the accuracy of the competent endoscopist was 0·956 (0·949–0·963; p<0·0001) and that of the trainee endoscopist was 0·886 (0·875–0·897; p<0·0001). The specificity was greater than 0·90 for all the three categories of endoscopists and GRAIDS. By contrast, sensitivity varied greatly among the endoscopists, and GRAIDS achieved similar sensitivity to the expert endoscopist (0·942 [95% CI 0·924–0·957] *vs* 0·945 [0·927–0·959]; p=0·692) and demonstrated significantly higher sensitivity than the competent endoscopist (0·858 [0·832–0·880]; p<0·0001) and trainee endoscopist (0·722 [0·691–0·752]; p<0·0001).

The PPV of GRAIDS was 0·814 (95% CI 0·788–0·838), which was significantly lower than that of the expert

	Accuracy	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Images (n=4532)	
						False positive detections	False negative detections
GRAIDS	0·928 (0·919–0·937)	0·942 (0·924–0·957)	0·923 (0·912–0·933)	0·814 (0·788–0·838)	0·978 (0·971–0·984)	237 (5·2%)	64 (1·4%)
Expert endoscopist	0·967 (0·961–0·973)	0·945 (0·927–0·959)	0·975 (0·968–0·981)	0·932 (0·913–0·948)	0·980 (0·974–0·985)	76 (1·7%)	61 (1·3%)
Competent endoscopist	0·956 (0·949–0·963)	0·858 (0·832–0·880)	0·992 (0·987–0·99)	0·974 (0·960–0·984)	0·951 (0·942–0·959)	25 (0·6%)	157 (3·5%)
Trainee endoscopist	0·886 (0·875–0·897)	0·722 (0·691–0·752)	0·945 (0·935–0·954)	0·824 (0·795–0·850)	0·904 (0·893–0·916)	175 (3·9%)	306 (6·8%)
GRAIDS and expert endoscopist	0·928 (0·919–0·937)	0·984 (0·973–0·991)	0·908 (0·896–0·920)	0·793 (0·768–0·818)	0·994 (0·989–0·996)	282 (6·2%)	18 (0·4%)
GRAIDS and competent endoscopist	0·934 (0·925–0·943)	0·978 (0·966–0·987)	0·919 (0·907–0·929)	0·812 (0·786–0·835)	0·992 (0·987–0·995)	250 (5·5%)	24 (0·5%)
GRAIDS and trainee endoscopist	0·904 (0·894–0·914)	0·964 (0·949–0·975)	0·883 (0·869–0·896)	0·747 (0·720–0·772)	0·985 (0·980–0·990)	360 (7·9%)	40 (0·9%)

Data are n (%) unless otherwise stated. GRAIDS=Gastrointestinal Artificial Intelligence Diagnosis System.

Table 3: **Performance of GRAIDS versus human endoscopists in identifying upper gastrointestinal cancers in a randomly selected subset of patients (n=175) from the prospective validation group**

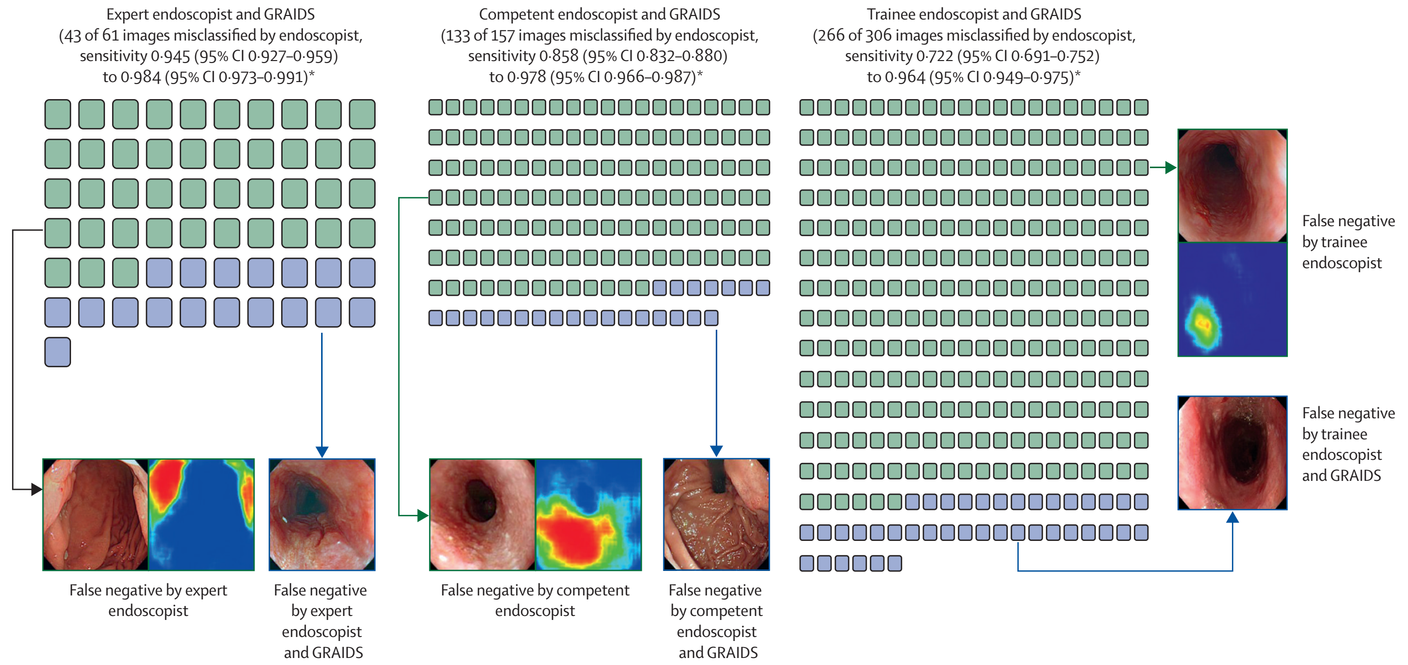
Expert endoscopist and GRAIDS
(43 of 61 images misclassified by endoscopist,
sensitivity 0·945 (95% CI 0·927–0·959)
to 0·984 (95% CI 0·973–0·991)*
Competent endoscopist and GRAIDS
(133 of 157 images misclassified by endoscopist,
sensitivity 0·858 (95% CI 0·832–0·880)
to 0·978 (95% CI 0·966–0·987)*
Trainee endoscopist and GRAIDS
(266 of 306 images misclassified by endoscopist,
sensitivity 0·722 (95% CI 0·691–0·752)
to 0·964 (95% CI 0·949–0·975)*
False negative by trainee endoscopist
False negative by trainee endoscopist and GRAIDS
False negative by expert endoscopist
False negative by expert endoscopist and GRAIDS
False negative by competent endoscopist
False negative by competent endoscopist and GRAIDS

Figure 3: **Performance of GRAIDS compared with human endoscopists in identifying upper gastrointestinal cancers in test images from a randomly selected subset of patients (n=175) from the prospective validation group**

The blue boxes refer to malignancies missed by both the endoscopists and GRAIDS, whereas the green ones indicate missed malignancies by the endoscopists but were successfully identified by GRAIDS. GRAIDS=Gastrointestinal Artificial Intelligence Diagnosis System. *p<0·0001.

endoscopist (0·932 [0·913–0·948]; p<0·0001) and the competent endoscopist (0·974 [0·960–0·984]; p<0·0001), but similar to the trainee endoscopist (0·824 [0·795–0·850]; p=0·580). When combined with GRAIDS, the expert, competent, and trainee endoscopists' PPVs all decreased significantly (0·793 [95% CI 0·768–0·818], 0·812 [0·786–0·835], and 0·747 [0·720–0·772], respectively; all p<0·0001). The NPV was high for GRAIDS (0·978 [95% CI 0·971–0·984]), endoscopists (expert, 0·980 [0·974–0·985]; competent, 0·951 [0·942–0·959]; and trainee, 0·904 [0·893–0·916]), and their combinations (table 3).

GRAIDS, however, was able to identify the majority of cancer images that were misclassified by the endoscopists (43 [70·5%] of 61 for the expert endoscopist, 133 [84·7%] of 157 for the competent endoscopist, and 266 [86·9%] of 306 for the trainee endoscopist; figure 3). When combined with GRAIDS, a significant but numerically modest improvement in sensitivity was observed for the expert (0·984 [95% CI 0·973–0·991], p<0·0001), and the sensitivity of the competent and trainee endoscopists significantly improved to a level similar to that of the expert (0·978 [0·966–0·987], p<0·0001, for the competent endoscopist and 0·964 [0·949–0·975], p<0·0001, for the trainee endoscopist).

The developed GRAIDS algorithm was capable of analysing as many as 118 images per second (8 ms per image) and processing a minimum of 25 images per second with a latency of less than 40 ms during real-time video analysis. In addition, we developed a computer-aided detection (CAD) system in an attempt for real-time identification of upper gastrointestinal cancerous lesions for use in routine endoscopic examination (appendix pp 9, 13). The computer on which the CAD system was installed was connected directly onto an endoscopy unit, thereby allowing fully automated diagnostic assistance during the endoscopic examinations.

Figure S5B (appendix p 9) and videos 1–4 (appendix p 13) demonstrate examples of the CAD system in real-time identification of cancerous lesions during an endoscopic examination. As illustrated, when GRAIDS identifies a malignant lesion, the CAD system segments the border of the lesion, as shown in blue, and warns the endoscopist about the possibility of a malignant lesion in the upper right corner of the screen. As the lesion disappears from the screen, the segmentation and warning signal simultaneously stop.

A cloud-based multi-institutional artificial intelligence platform (appendix p 10) was also constructed for patients requiring upper gastrointestinal endoscopy. This platform provides two key clinical applications: first, the real-time detection of upper gastrointestinal cancers during endoscopic procedures to aid in accelerating imaging interpretations and to assist in improving the accuracy of malignant lesion recognition. Second, storage of still images so they can be accessed post-examination to reassess suspicious cases, thus helping to decrease the risk of undetected and mis-detected malignancies.

For the **GRAIDS website** see http://graids.sysucc.org.cn/

A website has been made available to provide free access to GRAIDS (appendix p 10). Clinicians and patients can upload endoscopic images as a second-opinion consulting service for GRAIDS to review. An open-access endoscopic image database has also been made available on the website, which might be a useful resource for training endoscopists and to researchers in the field of endoscopy and artificial intelligence-aided medical imaging.

Discussion

In this study, we used a deep learning semantic segmentation model to construct an artificial intelligence-based upper gastrointestinal cancer diagnostic system, known as GRAIDS, that was trained and validated using 1 036 496 endoscopy images from 84 424 individuals, across six hospitals with different experiences and volumes of patients with an upper gastrointestinal cancer endoscopic diagnosis. GRAIDS demonstrated high accuracy, sensitivity, and specificity in detecting upper gastrointestinal cancers in retrospectively stored images as well as in a prospective observational setting. To the best of our knowledge, this is the largest study in the field of artificial intelligence-guided cancer detection based on upper gastrointestinal endoscopic images worldwide.

The endoscopic diagnosis of upper gastrointestinal cancer is subjective and to a great extent relies on the skills and experience of the physician.[11,21] Narrow-band imaging,[22] confocal laser endomicroscopy,[23] and blue laser imaging[24] have shown promising potential for differentiating between cancerous and non-cancerous lesions but their clinical applicability has been jeopardised because of the intensive training and expertise needed for optical image interpretation. By contrast, GRAIDS does not require additional training and has instead been found to improve the performance of non-expert endoscopists (competent from 0·858 to 0·978 and trainee from 0·722 to 0·964) to that approximating an expert level (0·967). Thus, for developing countries such as China or resource-limited countries, where there is an unbalanced distribution of medical resources between urban and rural areas, GRAIDS can help in bridging the cancer diagnosis gap between national hospitals and primary care hospitals.

The PPV of GRAIDS was found to be lower than that of expert and competent endoscopists, and the combination of GRAIDS with all three-level endoscopists resulted in decreased PPVs. In current real-time practice of endoscopic examination, GRAIDS reports suspicious cancer lesions without delineation contours marked by expert endoscopists, which might further increase the risk of false positives. However, the main causes of false positives by GRAIDS included normal structures or components such as the pylorus, gastric angle, and amount of mucus, as well as elevation of the gastric wall during peristalsis. Since these normal structures or changes would be easily recognised by endoscopists, misdiagnoses of these cases would probably be avoided in practice. Therefore, we speculate that in real-time endoscopy practice, when endoscopists are performing examinations using GRAIDS, the proportion of false positive cases would be lower than calculated. Moreover, the use of GRAIDS could lower the risk of missed diagnoses of cancer lesions as a result of its increased sensitivity, leading to earlier cancer detection, and could be considered as cost-effective in view of the high expenditure for treating upper gastrointestinal cancers.

The clinical applicability and advancement of existing upper gastrointestinal endoscopy have stalled as a result of the retrospective nature, small sample sizes, single disease investigations, and single institutional research at similar tier hospitals.[25,26] By comparison, GRAIDS was developed and validated using a large cohort of more than 1 million images from different tier hospitals and has exhibited an overall high accuracy (0·915–0·977) for the detection of upper gastrointestinal cancers in six retrospective validation sets, which strongly suggests the generalisability of the system in various real-world scenarios. In addition, the short imaging latency of less than 40 ms for image evaluation also makes it more efficient than existing yet-to-be prospectively investigated models (118 images per second *vs* 41·4[25] and 48·9[26] images per second).

On the basis of the accuracy and efficiency of GRAIDS in detecting upper gastrointestinal cancers, we constructed a cloud-based multi-institutional artificial intelligence platform, to provide fast and accurate real-time assistance during endoscopic procedures, and posteriori imaging evaluation. We also built a user-friendly website to provide freely accessible telemedical assistance for both patients and clinicians to accelerate the interpretation of endoscopic images. As of July 19, 2017, the Chinese Central & Southern Cancer Alliance (CCSCA) was founded, which aims to eliminate the gap in cancer management between national hospitals and primary care hospitals. Currently, GRAIDS is being routinely used in the endoscopic clinical workflow with real-time evaluation at SYSUCC and its screening centre, and is soon to be implemented by the other cooperating hospitals of the CCSCA, providing free-access to artificial intelligence-aided upper gastro-intestinal cancer screening and diagnosis.

Despite these remarkable results, there are some inherent limitations of GRAIDS worth highlighting. First, only white-light images were used for this study because such images are commonly used in both routine practice and resource-limited regions. Second, the training and external validation sets were labelled retrospectively, which might have led to a certain level of selection bias but the prospective validation suggests that this limitation might not be prominent. Third, we did not use specific approaches to handle the images at different locations from the same video series, which could create some inherit bias. Still, GRAIDS exhibited satisfactory accuracy across the participating hospitals, thereby demonstrating the general applicability of this system. Fourth, only high-quality endoscopic images for the training and validation analyses were used to investigate the optimal diagnostic efficacy of GRAIDS. Fifth, in the context of its clinical applicability, GRAIDS was trained and validated using a large Chinese cohort and its efficacy in other populations is yet to be investigated.

In conclusion, we have developed an artificial intelligence-based system using diversified endoscopic images from hospitals of different tiers that was able to achieve high diagnostic accuracy for detecting upper gastrointestinal cancers with sensitivity similar to that of expert endoscopists and superior to that of non-experts. GRAIDS could support non-expert endoscopists by improving their diagnostic accuracy to a level similar to that of experts. Furthermore, GRAIDS can provide real-time and retrospective assistance for improving the effectiveness of upper gastrointestinal cancer diagnosis and screening.

Contributors

GX, LH, CL, R-hX contributed to the study concept and design. LH, LL, BJ, YD, CH, and BL contributed to acquisition of data. LH, YL, ZW, YJ, WT, WT, YD, QW, JH, D-wH, S-bL, Q-mC, C-mY, FZ, YH, H-yP, H-xC, and BC contributed to analysis and interpretation of data. R-hX, HL, GX, ZW, YJ, LH, SRS, CL, LL, BJ, YD, YL, BL, WT, CH, QW, JH, D-wH, BC, S-bL, Q-mC, C-mY, H-xC, H-yP, FZ, and YH contributed to writing, reviewing, and approval of the final version of this work.

Declaration of interests

We declare no competing interests.

Data sharing

The stored still images from the Sun Yat-sen University Cancer Center are available on our website (http://graids.sysucc.org.cn/). Other datasets used or analysed during the current study are available from the corresponding author on reasonable request.

Acknowledgments

This research was supported by the National Key R&D Program of China (2018YFC1313300), the Natural Science Foundation of Guangdong Province (2017A030313485, 2014A030312015), the Science and Technology Program of Guangdong (2015B020232008), the Science and Technology Program of Guangzhou (201508020250, 201604020003, 2019B020227002), and the Fundamental Research Funds for the Central Universities (17ykpy82). We thank Christopher Lavender for providing assistance in editing this manuscript.

References

1 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA Cancer J Clin* 2018; **68:** 7–30.

2 Otutaha B, Srinivasa S, Koea J. Patient information needs in upper gastrointestinal cancer: what patients and their families want to know. *ANZ J Surg* 2019; **89:** 20–24.

3 Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin* 2017; **67:** 93–99.

4 Sano T, Coit DG, Kim HH, et al. Proposal of a new stage grouping of gastric cancer for TNM classification: International Gastric Cancer Association staging project. *Gastric Cancer* 2017; **20:** 217–25.

5 Rice TW, Ishwaran H, Hofstetter WL, et al. Recommendations for pathologic staging (pTNM) of cancer of the esophagus and esophagogastric junction for the 8th edition AJCC/UICC staging manuals. *Dis Esophagus* 2016; **29:** 897–905.

6 Veitch AM, Uedo N, Yao K, East JE. Optimizing early upper gastrointestinal cancer detection at endoscopy. *Nat Rev Gastroenterol Hepatol* 2015; **12:** 660–67.

7 Chiu PWY, Uedo N, Singh R, et al. An Asian consensus on standards of diagnostic upper endoscopy for neoplasia. *Gut* 2019; **68:** 186–97.

8 Hamashima C, Systematic Review Group and Guideline Development Group for Gastric Cancer Screening Guidelines. Update version of the Japanese Guidelines for gastric cancer screening. *Jpn J Clin Oncol* 2018; **48:** 673–83.

9 Jun JK, Choi KS, Lee HY, et al. Effectiveness of the Korean National Cancer Screening Program in Reducing Gastric Cancer Mortality. *Gastroenterology* 2017; **152:** 1319–28.e7.

10 Wong Kee Song LM, Wilson BC. Endoscopic detection of early upper GI cancers. *Best Pract Res Clin Gastroenterol* 2005; **19:** 833–56.

11 Menon S, Trudgill N. How commonly is upper gastrointestinal cancer missed at endoscopy? A meta-analysis. *Endosc Int Open* 2014; **2:** e46–50.

12 Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol* 2019; **20:** 193–201.

13 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–18.

14 Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med* 2019; **25:** 433–38.

15 Trister AD, Buist DSM, Lee CI. Will machine learning tip the balance in breast cancer screening? *JAMA Oncol* 2017; **3:** 1463–64.

16 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.

17 Hirasawa T, Aoyama K, Tanimoto T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer* 2018; **21:** 653–60.

18 Alagappan M, Brown JRG, Mori Y, Berzin TM. Artificial intelligence in gastrointestinal endoscopy: the future is almost here. *World J Gastrointest Endosc* 2018; **10:** 239–49.

19 Pohlen T, Hermans A, Mathias M, Leibe B. Full-resolution residual networks for semantic segmentation in street scenes. arXiv 2016; published online Dec 6. https://arxiv.org/abs/1611.08323 (preprint).

20 Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv 2018; published online Aug 22. https://arxiv.org/abs/1802.02611 (preprint).

21 Yamazato T, Oyama T, Yoshida T, et al. Two years' intensive training in endoscopic diagnosis facilitates detection of early gastric cancer. *Intern Med* 2012; **51:** 1461–65.

22 Muto M, Minashi K, Yano T, et al. Early detection of superficial squamous cell carcinoma in the head and neck region and esophagus by narrow band imaging: a multicenter randomized controlled trial. *J Clin Oncol* 2010; **28:** 1566–72.

23 Neumann H, Kiesslich R, Wallace MB, Neurath MF. Confocal laser endomicroscopy: technical advances and clinical applications. *Gastroenterology* 2010; **139:** 388–92.

24 Dohi O, Yagi N, Naito Y, et al. Blue laser imaging-bright improves the real-time detection rate of early gastric cancer: a randomized controlled study. *Gastrointest Endosc* 2019; **89:** 47–57.

25 Horie Y, Yoshio T, Aoyama K, et al. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointest Endosc* 2019; **89:** 25–32.

26 Hirasawa T, Aoyama K, Tanimoto T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer* 2018; **21:** 653–60.