OPEN ACCESS

ORIGINAL ARTICLE

Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy

Lianlian Wu,[1,2,3] Jun Zhang,[1,2,3] Wei Zhou,[1,2,3] Ping An,[1,2,3] Lei Shen,[1,2,3] Jun Liu,[1,3] Xiaoda Jiang,[1,2,3] Xu Huang,[1,2,3] Ganggang Mu,[1,2,3] Xinyue Wan,[1,2,3] Xiaoguang Lv,[1,2,3] Juan Gao,[1,3] Ning Cui,[1,2,3] Shan Hu,[4] Yiyun Chen,[4] Xiao Hu,[4] Jiangjie Li,[4] Di Chen,[1,2,3] Dexin Gong,[1,2,3] Xinqi He,[1,2,3] Qianshan Ding,[1,2,3] Xiaoyun Zhu,[1,2,3] Suqin Li,[1,2,3] Xiao Wei,[1,2,3] Xia Li,[1,2,3] Xuemei Wang,[1,2,3] Jie Zhou,[1,2,3] Mengjiao Zhang,[1,2,3] Hong Gang Yu[1,2,3]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2018-317366).

For numbered affiliations see end of article.

Correspondence to
Professor Hong Gang Yu, Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan 430060, China; yuhonggang1968@163.com

LW, JZ and WZ contributed equally.

Received 10 August 2018
Revised 28 January 2019
Accepted 17 February 2019

Check for updates

© Author(s) (or their employer(s)) 2019. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Wu L, Zhang J, Zhou W, *et al*. *Gut* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/gutjnl-2018-317366

ABSTRACT

Objective Esophagogastroduodenoscopy (EGD) is the pivotal procedure in the diagnosis of upper gastrointestinal lesions. However, there are significant variations in EGD performance among endoscopists, impairing the discovery rate of gastric cancers and precursor lesions. The aim of this study was to construct a real-time quality improving system, WISENSE, to monitor blind spots, time the procedure and automatically generate photodocumentation during EGD and thus raise the quality of everyday endoscopy.

Design WISENSE system was developed using the methods of deep convolutional neural networks and deep reinforcement learning. Patients referred because of health examination, symptoms, surveillance were recruited from Renmin hospital of Wuhan University. Enrolled patients were randomly assigned to groups that underwent EGD with or without the assistance of WISENSE. The primary end point was to ascertain if there was a difference in the rate of blind spots between WISENSE-assisted group and the control group.

Results WISENSE monitored blind spots with an accuracy of 90.40% in real EGD videos. A total of 324 patients were recruited and randomised. 153 and 150 patients were analysed in the WISENSE and control group, respectively. Blind spot rate was lower in WISENSE group compared with the control (5.86% vs 22.46%, p<0.001), and the mean difference was −15.39% (95% CI −19.23 to −11.54). There was no significant adverse event.

Conclusions WISENSE significantly reduced blind spot rate of EGD procedure and could be used to improve the quality of everyday endoscopy.

Trial registration number ChiCTR1800014809; Results.

Significance of this study

What is already known on this subject?

► The past decades have seen remarkable progress of deep convolutional neural network (DCNN) in the field of endoscopy. Recent studies have successfully used DCNN to achieve accurate prediction of early gastric cancer in endoscopic images and real-time histological classification of colon polyps in unprocessed videos. However, it has yet not been investigated whether DCNN could be used in monitoring quality of everyday endoscopy.

What are the new findings?

► In the present study, WISENSE, a real-time quality improving system based on the DCNN and deep reinforcement learning (DRL) for monitoring blind spots, timing the procedure and generating photodocumentation during esophagogastroduodenoscopy (EGD) was developed. The performance of WISENSE was verified in EGD videos. A single-centre randomised controlled trial was conducted to evaluate the hypothesis that WISENSE would reduce the rate of blind spots during EGD. To the best of our knowledge, this is the first study using deep learning in the field of assuring endoscopy completeness and using DRL in making medical decisions in human body environment and also the first study validating the efficiency of a deep learning system in a randomised controlled trial.

How might it impact on clinical practice in the foreseeable future?

► WISENSE greatly reduced blind spot rate, increased inspection time and improved the completeness of photodocumentation of EGD in the randomised controlled trial. It could be a powerful assistant tool for mitigating skill variations among endoscopists and improving the quality of everyday endoscopy.

INTRODUCTION

Esophagogastroduodenoscopy (EGD) is the pivotal procedure in the diagnosis of upper gastrointestinal lesions.[1] High-quality endoscopy delivers better health outcomes.[2] However, there are significant variations in EGD performance among endoscopists, impairing the discovery rate of gastric cancers (GC) and precursor lesions.[3] The diagnosis rate of early GC in China is still

Gut: first published as 10.1136/gutjnl-2018-317366 on 11 March 2019. Downloaded from http://gut.bmj.com/ on 13 March 2019 by guest. Protected by copyright.

under 20% and similar results are seen in most part of the world.[4 5] While further expanding endoscopic technology, it is vital to raise the quality of everyday endoscopy.

Ensuring competence is a seminal prerequisite for discovering lesions in EGD.[6] Plenty of guidelines or expert consensus have been reached to optimise EGD examination.[7] The American Society for Gastrointestinal Endoscopy (ASGE) and American College of Gastroenterology (ACG) developed safety and quality indicators for EGD.[8–10] In 2015, European Society of Gastrointestinal Endoscopy (ESGE) systematically investigated available evidences and generated the first evidence-based performance measures of EGD.[1] Standardised protocols were proposed to fully map the entire stomach.[11 12] However, protocols are not often well followed due to the shortage of supervision and the availability of practical tools, especially in developing countries.[2] It is essential to establish practical and feasible methods to implement guidelines in daily endoscopy.

The past decades have seen remarkable progress of deep learning in medicine.[13] Deep convolutional neural network (DCNN) is known for its impressive performance in image recognition.[14] Recent studies have successfully used DCNN to achieve accurate prediction of early GC in endoscopic images[15] and real-time histological classification of colon polyps in unprocessed videos.[16] However, previous researches mainly focus on the diagnosis of gastrointestinal lesions. It has yet not been investigated whether DCNN could be used in monitoring quality of everyday endoscopy.

In addition, in real clinical setting, doctors always make comprehensive judgments based on more than one frame due to dynamic and constantly changing views in human body. While DCNN analyses frames independently,[14] and there are plenty of noises in real world.[17] This may cripple the application of DCNN in real clinical setting. Deep reinforcement learning (DRL), another branch of deep learning that won its recent reputation in the game of Go in 2016,[18] may have the potential to solve this problem. DRL is a combination of deep learning and reinforcement learning, integrating the strong perception ability of deep learning in visual tasks, and the decision-making ability of reinforcement learning in complex situations.[19] DRL shows good performance in solving dynamic decision problems;[18–20] however, it has yet not been explored in making medical decisions in human body environment.

In the present study, we aimed to construct a real-time quality improving system based on the two methods of DRL and DCNN for monitoring blind spots, timing the procedure and generating photodocumentation during EGD, which was named WISENSE (a combination of 'wise' and 'sense'). The performance of WISENSE was verified in EGD videos. A single-centre randomised controlled trial was conducted to evaluate the hypothesis that WISENSE would reduce the rate of blind spots during EGD. To the best of our knowledge, this is the first study using deep learning in the field of assuring endoscopy completeness and using DRL in making medical decisions in human body environment and also the first study validating the efficiency of a deep learning system in a randomised controlled trial.

MATERIALS AND METHODS

Datasets and preprocessing

Three datasets were used for training and/or testing the WISENSE:

1. 12 220 in vitro, 25 222 in vivo and 16 760 unqualified EGD images for training the network to identify whether a scope was in or outside the body (DCNN1). These images came from stored data of over 3000 patients. Two doctorial students labelled these images and their labels were combined by consensus. Representative images of DCNN1 were shown in online supplementary figure S1.
2. 34 513 qualified EGD images for training the network of classifying gastric sites (DCNN2). Two seniors with 1–5 years of EGD experience and three experts with more than 5 years of EGD experience studied the guidelines of the ESGE[1] and the Japanese systematic screening protocol[11] and independently labelled EGD images into 26 different sites or NA. To alleviate the incorporated bias of single endoscopist, images were labelled only when no less than four endoscopists reached an agreement. Images including features such as forceps and lesions were also included in each site to prevent the system from associating the appearance of tools or lesions with different sites. Representative images of DCNN2 were shown in figure 1. Sample distribution in different classifications is presented in online supplementary table S1.
3. 30 stored EGD videos were used to identify the best status of DRL. A total of 107 stored EGD videos were used to test the performance of WISENSE in clinical setting.

One-tenth of labelled images in each classification of dataset (1) and (2) were extracted as the testing set, with the remaining as training set. Extensive attention was paid to ensure that images from the same person were not split between the training and testing sets. All the EGD images and videos were in white light view and from Renmin Hospital of Wuhan University. Instruments used in this study included gastroscopes from two vendors (Olympus Optical Co., Tokyo, Japan; Fujifilm, Co., Kanagawa, Japan).

Main experiments

1. Training and testing of DCNN in still images.
2. Traini ng of DRL on virtual EGD videos and testing on 30 real EGD videos.
3. Integrating DCNN and DRL, and testing on the 107 real EGD videos.
4. Conducting a single-centre randomised controlled trial to evaluate the effect of WISENSE on the quality of EGD examination.

Training and testing of DCNN

VGG-16 and DenseNet, two mature DCNN models pretrained with 1.28 million images from 1000 object classes,[21] were used to train our system. K-fold cross-validation procedure[22] was implemented with k=10, dividing the training dataset into 10 subsets and validating each subset individually with the remaining used for training. Detailed technical methods and methods of avoiding overfitting were described in supplementary methods and materials.

In the testing sets, DCNNs were tested for three times, with each time based on randomly selected 3000 images (1000 per category) from the testing set of DCNN1 and 2160 images (80 per site) from the testing set of DCNN2. Online supplementary figure S2 shows confusion matrices indicating the performance of VGG-16 and DenseNet on classifying each category. VGG-16 and DenseNet achieved a comparable accuracy (97.55%±0.18% and 97.86%±0.19%, respectively) in DCNN1, while VGG-16 showed a superior accuracy to that of DenseNet (88.70%±0.23% and 83.76%±0.22%, respectively) in DCNN2. Therefore, VGG-16 was chosen to further develop WISENSE.

Gut: first published as 10.1136/gutjnl-2018-317366 on 11 March 2019. Downloaded from http://gut.bmj.com/ on 13 March 2019 by guest. Protected by copyright.

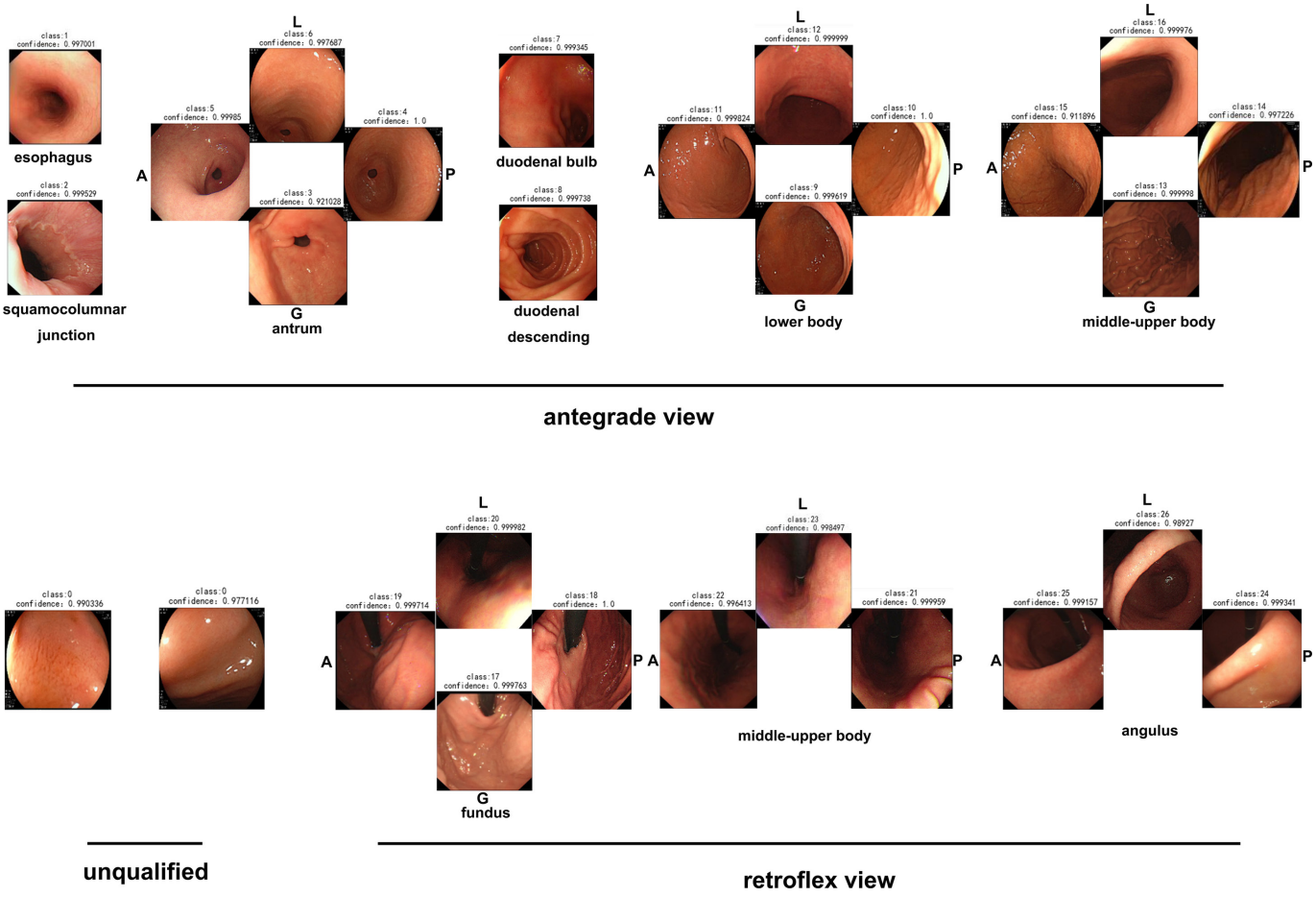

Figure 1 Representative images predicted by the WISENSE in classifying gastric images into 26 sites or NA. The displays showed the gastric sites determined by the WISENSE and the prediction confidence. Class 0, NA, images that could not be classified in any site due to the absence of anatomical landmarks. (1) oesophagus; (2) squamocolumnar junction; (3–6) antrum (G, P, A, L); (7) duodenal bulb; (8) duodenal descending; (9–12) lower body (G, P, A, L); (13–16) middle-upper body in forward view (G, P, A, L); (17–20) fundus (G, P, A, L); (21–23) middle-upper body in retroflex view (P, A, L); (24–26) angulus (P, A, L). A, anterior wall; G, greater curvature; L, lesser curvature; P, posterior wall.

Training and testing of DRL

To make human logicality for predicting gastric sites and reduce noise signals in real EGD videos, an agent based on DRL was conducted. DRL is a way originated from behavioural psychology, where agents learn tasks by being given a reward for a correct action and a punishment for a wrong action in given states and creating a self-learning feedback loop.[20] Using neuroscience as an analogy, we used a reward/penalty mechanism (representing dopamine) to train a DRL model (representing the prefrontal cortex) that could take actions (decisions made by the prefrontal cortex) in different states (the environment where we are located).[23] In a typical DRL task, a state generally refers to a snapshot of everything in an environment at a certain time.[20] Here, we defined a state as labels and confidences of the previous nine consecutive images predicted by the DCNN, and all gastric sites previously activated by DRL at a certain time. This information was projected into a 10×26 matrix that can be input to DRL model (figure 2), as explained in detail in online supplementary methods and materials. DRL will make an action based on the state, lighting a site of 1–26 or keeping silence (score 0) and get a reward for a correct action (score +3) or a penalty for a wrong action (score −6). At the beginning of the training, the model randomly takes actions, and as the training goes on, more and more experiences are gained and stored in an experience pool.[24] Then the model could randomly extract previous experiences from the experience pool and take an optimal action in a state using the basic reinforcement algorithm (Bellman equation).[25]

$$Q(s, a) = r + \gamma \max a' Q(s', a')$$

It was designed to achieve a maximum cumulative reward [Q(s,a)] with both the immediate reward (r) and maximum future reward for the next state [maxa′Q(s′,a′)] taken into consideration. The process the DRL chooses how to take action, just like the process how AlphaGo decides to play in the Go game. In this equation, r is the benefit obtained from the current step (immediate reward). Maxa'Q(s',a') is the assumed maximum benefit in the next step (future reward), inferred by the neural network through previous experiences and constantly changing with every step. γ is a discount factor multiplying maxa'Q(s',a'), representing the model's emphasis degree on future rewards. We fitted EGD video data to the model to estimate the necessary parameter and set the γ=0.2, which indicates that 20% expected future rewards are considered when taking actions.

Virtual EGD videos were randomly generated based on basic principles of EGD procedure (entering from the oesophagus into the gastrointestinal and then exit the oesophagus). In every epoch, DRL was trained on 50 virtual EGD videos and then tested on 30 real EGD videos. As shown in online supplementary figure S3, after 74 epochs, the DRL model converged and

Gut: first published as 10.1136/gutjnl-2018-317366 on 11 March 2019. Downloaded from http://gut.bmj.com/ on 13 March 2019 by guest. Protected by copyright.

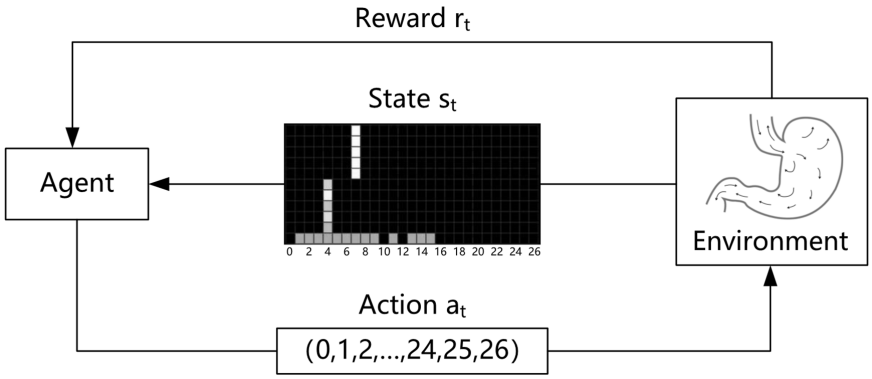

Figure 2 A diagram of the DRL model. DRL makes an action (a_t) based on the state (s_t) in environment, lighting a site of 1–26 or do nothing (action is 0) and get a reward (positive score) for a correct action. Labels and confidences of images are projected into a 10×26 grid into a state that can be input to the DRL. Numbers in the abscissa of the matrix represents 26 gastric sites or NA, and the ordinate represents when frames appear. Small cubes in the nine rows from top to bottom represent EGD frames appeared in different times, with their respective positions in abscissa showing their sites predicted by DCNN. The colour shade of cubes represents the confidence of the DCNN's prediction (the whiter, the higher). The cube representing the first frame appears at the top of the matrix when a video is played, and the previous cube moves down and the next cube appears at the top when the second frame comes. Cubes keep falling down from top to bottom, and for a while, we could see nine cubes dynamically displayed in the matrix until the end of the video, showing predictions and confidences of DCNN on nine consecutive frames. Grey cubes in the bottom row of the matrix show sites that identified to be observed by DRL. DCNN, deep convolutional neural networks; DRL, deep reinforcement learning; EGD, esophagogastroduodenoscopy.

achieved an accuracy of 91.40%. To make the model robust enough, DRL was trained using DCNN predictions with a 20% false probability, and this led to a course that the testing accuracy is higher than training accuracy.

After training of DCNN and DRL, we tested and compared the performance of WISENSE combining DCNN with DRL to a system combining DCNN with a traditional method (random forest filtering), as described in online supplementary methods and materials in detail. WISENSE and the system were set to process images at 2 frames per second (fps) on videos.

A single-centre randomised controlled trial

Trail design

This was a prospective, single-centre, single-blind, randomised, parallel-group study, approved by the Ethics Committee of Renmin Hospital of Wuhan University and under trail registration number ChiCTR1800014809 of the Primary Registries of the WHO Registry Network.

Participants

Consecutive patients undergoing routine EGD examinations in the endoscopy centre of Renmin Hospital, Wuhan University, China between August and October 2018 were recruited to the study. Patients were followed up until they wake up after EGD.

Inclusion criteria: (1) patients aged 18 years or older; (2) American Society of Anesthesiology risk class 1, 2 or 3; (3) patients able to give informed consent.

Exclusion criteria: (1) patients with absolute contraindications to EGD examination; (2) a history of previous gastric surgery; (3) patients in pregnancy; (4) allergic to anaesthetic in previous medical history; (5) researchers believe that the patient is not suitable to participate in the trial.

Withdrawal criteria: (1) the EGD procedure cannot be completely conducted due to oesophageal stenosis, obstruction, huge occupying lesions or giant ulcer of the duodenal bulb; (2) gastroscope must be withdrawn in advance due to rapid changes in patients' heart rate or breathing rate.

The population to patients were not limited to particular indicators, because most patients with early-stage GC are asymptomatic,[26 27] and WISENSE was developed based on the 'systematic screening protocol for the stomach (SSS)', which was proposed as a minimum required standard for routine endoscopy.[11]

Enrolled endoscopists were six staff members in the Gastroenterology Department of Renmin Hospital, Wuhan University, with EGD experience of 1–3 years, and EGD volume was 2000–5000. The six endoscopists studied the working interface of WISENSE and the standard protocol of the 26 sites before the trial.

Interventions

Patients were randomly assigned to undergo an EGD examination with or without the assistance of WISENSE. In the WISENSE group, except for the original videos, there were three additional information presented to endoscopists: (1) the virtual stomach model monitoring blind spots; (2) timing; (3) scoring and grading: to increase the intuition of WISENSE, the number of observed sites was converted in percentage term and graded to 'good', 'excellent' and 'perfect' when the score reached 80, 90 and 100, respectively. In addition, WISENSE automatically extracted and collected frames with the highest confidence in each site during EGD (figure 3). In the control group, there was no additional information presented and no photodocumentation collected. Figure 4 and online video 1 shows the real-time working of WISENSE system.

Outcomes

The primary outcome of the study was the blind spot rate (number of unobserved sites in each patient/26) in WISENSE and control groups.

The secondary outcomes were: (1) inspection time; (2) completeness of photodocumentation generated by endoscopists; (3) completeness of photodocumentation generated by WISENSE in WISENSE group; (4) completeness of photodocumentation generated by WISENSE and endoscopists in

Gut: first published as 10.1136/gutjnl-2018-317366 on 11 March 2019. Downloaded from http://gut.bmj.com/ on 13 March 2019 by guest. Protected by copyright.

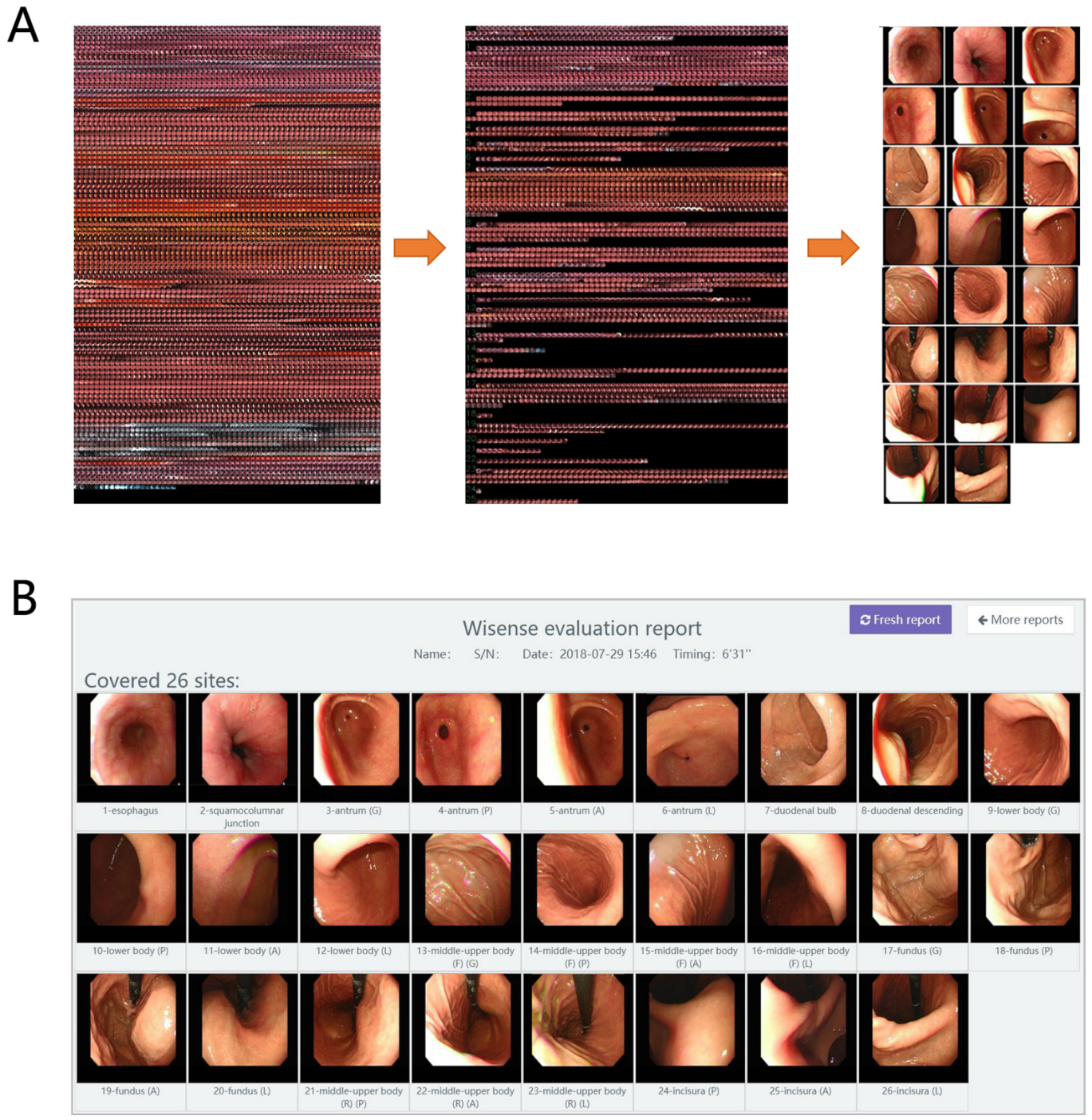

Figure 3 A schematic illustration of how the WISENSE obtains photodocumentation during EGD procedure. (A) For obtaining accurate photodocumentation, WISENSE first filtered unqualified images and then extracted the most representative frame in each site during the process of EGD. (B) A representative photodocumentation generated by WISENSE. A, anterior wall; G, greater curvature; F, forward view; L, lesser curvature; P, posterior wall; R, retroflex view. EGD, esophagogastroduodenoscopy.

WISENSE group; (5) the per cent of patients being ignored in each site.

Two seniors with 1–5 years of EGD experience and three experts with more than 5 years of EGD experience independently reviewed the videos from the RCT to document blind spots, start time and end time of each case and reviewed photodocumentation generated by WISENSE or endoscopists to document covered sites. Their results were combined by consensus. Endoscopists performing EGD examinations did not participate in data evaluation. Video recording and storage were described in online supplementary methods and materials.

Sample size

The study was powered to demonstrate a difference in blind spot rate during EGD with or without the assistance of WISENSE. In previous studies, there are rare data analysing blind spots during EGD procedure. Therefore, we conducted a pilot study and estimated that mean blind spot rate was 0.20 in control and 0.10

Gut: first published as 10.1136/gutjnl-2018-317366 on 11 March 2019. Downloaded from http://gut.bmj.com/ on 13 March 2019 by guest. Protected by copyright.

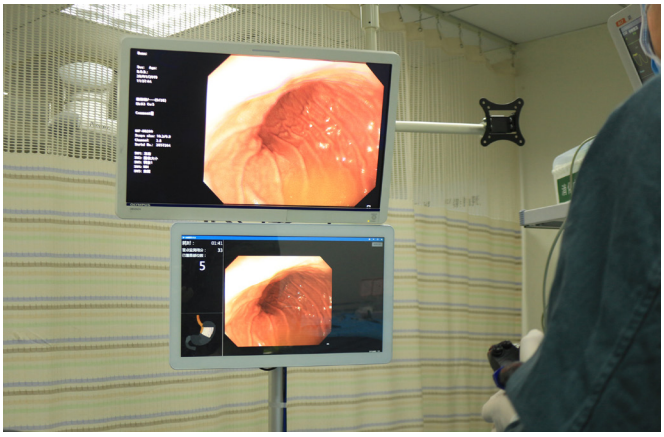

Figure 4 Real-time use of WISENSE with an endocytoscope during esophagogastroduodenoscopy. The computer on which WISENSE is installed was directly connected to an endoscopy unit (Evis Lucera Elite CV 290, Olympus) and placed side by side with the original screen, achieving real-time monitoring blind spots during esophagogastroduodenoscopy.

in WISENSE-assisted EGD, and the SD was 0.15. With a power of 90%, a two-sided significance level of 0.05 and a superiority margin of 0.05, 154 patients per group were required. Assuming that approximately 5% of patients may be excluded from the analysis, the target sample size was set as 162 per group. Notably, the rate of blind spot is a discrete continuous variable, and the sample size was calculated using the method of comparing two means[28] with online power calculators (http://powerandsamplesize.com).

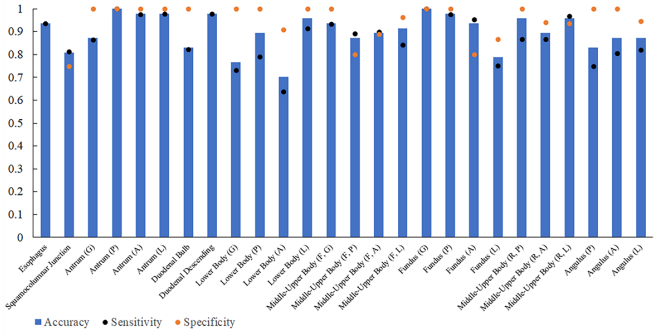

Figure 5 The accuracy, sensitivity and specificity of WISENSE for monitoring blind spots in real EGD videos. In 107 real EGD videos, WISENSE monitored blind spots with an average accuracy of 90.02%, and a separate accuracy for each site ranging from 70.21% to 100%. The average sensitivity and specificity were of 87.57% and 95.02%, ranging from 63.4% to 100% and 75% to 100%, respectively. All EGD videos contain the oesophagus and duodenum; therefore, the negative value of oesophagus and duodenum was zero and specificity of the two sites was unavailable. True positive, WISENSE lights up site A in the stomach model when endoscopists also label site A; true negative, WISENSE leaves site B in transparent in the stomach model and site B is also not labelled by endoscopists. The number of videos containing site C is the 'positive' value of site C, and the number of videos missing site D is the 'negative' value of site D. Acccuracy=true predictions/(positive+negative), sensitivity=true positive/positive, specificity=true negative/negative. EGD, esophagogastroduodenoscopy.

Randomisation and blinding

The random allocation sequence was generated by computer-generated random numerical series with '0' encoding for WISENSE group and '1' encoding for control group. Randomisation was done in blocks of 4 in a 1:1 ratio. Endoscopists were not blinded to randomisation status, while patients and all study analyses were conducted in a blinded fashion.

Statistical analysis

χ^2 test was used to compare baseline characteristics and the per cent of patients being ignored in each site between groups. Mann Whitney U statistic with a significance level of 0.05 was used to compare the primary outcome and other secondary outcomes between groups. As for the primary outcome, a higher 95% CI bound of less than −0.05 of the difference for blind spot rate between two groups was required to confirm the superiority of WISENSE. All analyses were performed using StatsDirect V.3.1.20 (StatsDirect Ltd).

RESULTS

The performance of WISENSE in real EGD videos

WISENSE monitored blind spots with an average accuracy of 90.02%, and a separate accuracy for each site ranging from 70.21% to 100% in the 107 real EGD videos. The average sensitivity and specificity were of 87.57% and 95.02%, ranging from 63.4% to 100% and 75% to 100%, respectively (figure 5). Notably, all EGD videos contain the oesophagus and duodenum, therefore, the negative value of oesophagus and duodenum was zero and specificity of the two sites was unavailable. The performance of a system combining DCNN with a traditional method (random forest filtering) was tested and compared with WISENSE in online supplementary methods and materials.

For timing EGD procedure, WISENSE correctly predicted the start time in 93.46% (100/107) videos and end time in 97.20% (104/107) videos. Cases of early or delayed timing are described in supplementary methods and materials.

Recruitment

A total of 361 patients were invited to participate in the trial. Thirty-seven patients were excluded as they were ineligible (n=33) and declined participation (n=4). A total of 324 patients were recruited and randomised, and the trial flowchart is illustrated in figure 6. A total of 153 patients from the WISENSE group and 150 patients from the control group were included in the analysis. Patient characteristics were comparable in both groups (table 1).

Outcomes

Blind spot rate was significantly lower in WISENSE group compared with the control (5.86% vs 22.46%, p<0.001), and the mean difference was −15.39% (95% CI, −19.23 to −11.54). The higher 95% CI bound is less than −5% of the difference for blind spot rate between two groups (table 2).

Mean inspection of EGD procedure was significantly longer in WISENSE group compared with the control (5.03 min vs 4.24 min, p<0.001). Considering that there were still 5.86% sites being ignored in the WISENSE group, we further analysed the inspection time of those cases that all 26 sites were observed, and it turned out to be 5.36±2.97 min.

There was no difference in the completeness of photodocumentation generated by endoscopists between the two groups (71.87% vs 79.14%, p=0.11). However, the completeness of photodocumentation generated by WISENSE in WISENSE group

Gut: first published as 10.1136/gutjnl-2018-317366 on 11 March 2019. Downloaded from http://gut.bmj.com/ on 13 March 2019 by guest. Protected by copyright.

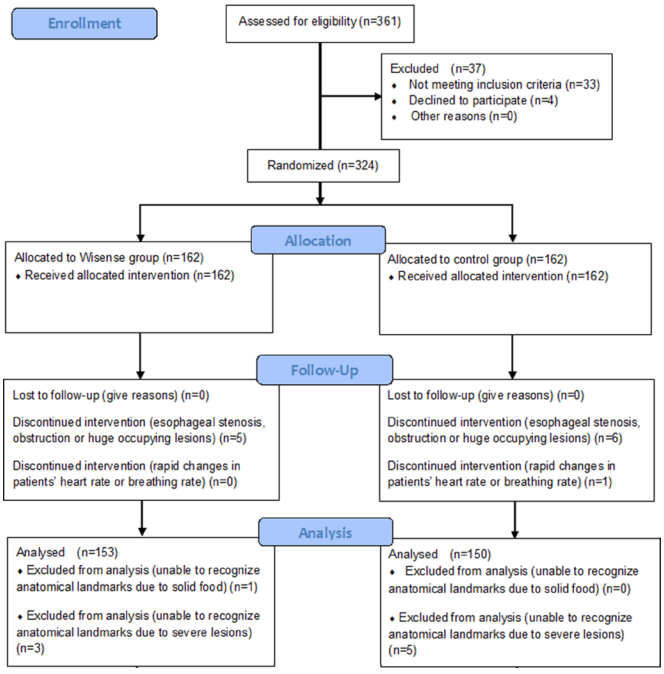

Figure 6 Trial flow diagram.

was significantly higher than that from endoscopists in control group (90.64% vs 79.14%, p<0.001). When images collected by WISENSE and endoscopists were combined, the completeness of photodocumentation further increased in WISENSE group and the difference between groups raised (92.91% vs 79.14%, p<0.001).

The per cent of patients being ignored in most gastric sites (80.77%, 21/26) were significantly lower in WISENSE group compared with that in control group (table 3). The per cent of patients being ignored in the lesser curvature of middle-upper body in forward view decreased most (−47.50% RD, p<0.001). No patient's oesophagus was ignored in both groups. No patient's squamocolumnar junction was ignored in WISENSE group while there were two patients' squamocolumnar junction that was ignored in control group. There was no difference of patients being ignored in the anterior wall of antrum, the greater curvature of middle-upper body in forward view and the anterior wall of fundus.

Table 1 Baseline characteristics

Characteristics	WISENSE (n=153)	Control (n=150)	P value
Age, mean (SD)	50.6 (14.2)	49.1 (13.4)	0.34
Female, n (%)	77 (50.3)	81 (54.0)	0.52
Indications for EGD, n			0.85
Abdominal discomfort	103	101	
Diarrhoea	3	2	
Health examination	21	24	
Acid reflux	9	8	
Suspected GI bleeding	3	5	
Bowel habit change	1	1	
Dyspepsia	4	2	
Belching	3	1	
Anaemia	1	0	
Constipation	1	0	
Vomiting	2	0	
Suspected malignancy	2	2	
Emaciation	0	2	
Poor appetites	0	1	
Dysphagia	0	1	
Recruitment, n (%)			0.72
Inpatient	41 (26.8)	43 (28.1)	
Outpatient	112 (73.2)	107 (69.9)	

EGD, esophagogastroduodenoscopy.

DISCUSSION

Hundreds of millions of people undergo EGD procedures every year worldwide. High-quality endoscopy delivers better health outcomes.[2] However, there are significant variations in EGD performance among endoscopists, impairing the discovery of upper GI lesions.[3] A bunch of guidelines have been proposed to standardise EGD;[7] however, they are not often well followed due to the shortage of supervision and availability of practical tools, especially in developing countries.[2] WISENSE, a real-time quality improving system for monitoring blind spots during EGD, was developed to specifically address this need. The system monitored blind spots in real EGD videos with an accuracy of 90.02%, played roles of real-time quality supervision and automatically generating photodocumentation. In a clinical trial, the WISENSE greatly reduced blind spot rate, increased inspection time and improved completeness of photodocumentation in EGD.

Recent years have seen an explosion of study in the application of DCNN in endoscopy. Byrne *et al* achieved real-time differentiation of adenomatous and hyperplastic colorectal polyps.[16] Hirasawa T *et al* used the DCNN to detect early GC with a sensitivity of 92.2%.[15] Most published work focus on the diagnosis or detection of lesions; however, there was rare research being conducted to mitigate technical variations of endoscopists. GC occurs in every part of the gastric cavity.[29] Observing the whole stomach is a seminal prerequisite for EGD examination.[1 11] In China, the majority of small (<2 cm) GC (68%) are concentrated in the lesser curvature, mainly in two areas, the fundus and antrum.[29] However, we found that there were 40.67% and 19.33% patients not being observed in control EGD in these two sites, respectively. With the assistance of WISENSE, patients not being observed in the lesser curvature of fundus and antrum were decreased to less than 20% and 4%, respectively. This indicates that WISENSE may increase the detection of early GC or precancerous lesions, and thus improve the health outcome of patients. To the best of our knowledge, this is the first study applying deep learning in field of supervising endoscopy quality and validating the efficiency of a deep learning system in a randomised controlled trial.

Accurate photodocumentation allows for better knowledge and follow-up of patients' condition.[1 9] Ninety-one per cent agreement was reached among the experts from ESGE for accurate photodocumentation in EGD.[1] In the present study, there was no difference in the completeness of photodocumentation generated by endoscopists with or without WISENSE (71.87% vs 79.14%). However, the completeness of photodocumentation generated by WISENSE (90.64%) was significantly higher than that from endoscopists. After combining photodocumentation of both endoscopists and WISENSE, the completeness further improved (92.91%). In daily clinical work, endoscopists need to pause operation or obtain help from an assistant to capture images. The WISENSE could automatically help endoscopists

Gut: first published as 10.1136/gutjnl-2018-317366 on 11 March 2019. Downloaded from http://gut.bmj.com/ on 13 March 2019 by guest. Protected by copyright.

Table 2 Primary and secondary outcomes for all patients

Endpoint	Mean (SD) WISENSE (n=153)	Control (n=150)	Difference (95% CI)	P value
Primary endpoint				
Blind spot rate	5.86 (6.89)	22.46 (14.38)	−15.39 (−19.23 to −11.54)	<0.001
Secondary endpoints				
Inspection time (min)	5.03 (2.95)	4.24 (3.82)	0.90 (0.43 to 1.35)	<0.001
Photodocumentation completeness (%)				
Endoscopists *vs* endoscopists	71.87 (29.43)	79.14 (21.89)	−3.85 (−9.09 to 0)	0.11
WISENSE *vs* endoscopists	90.64 (9.80)	79.14 (21.89)	7.11 (3.42 to 10.76)	<0.001
WISENSE and endoscopists *vs* endoscopists	92.91 (21.16)	79.14 (21.89)	11.77 (8.70 to 15.79)	<0.001

generate photodocumentation and improve the completeness of photodocumentation.

Another strength of this system is that it times EGD procedures in real time and record inspection time. Longer examination time spent on EGD improves the diagnostic yield.[30] The elapsed duration of EGD procedure was often recorded by endoscopy nurses via wall clocks in endoscopy rooms and records from nurses were not often actively reported to endoscopists.[30] WISENSE could automatically time during EGD and record inspection time for every procedure. This may help endoscopists to monitor and control the time spent on each procedure and thus mitigate skill variations from subjective factors and external pressure.

Noise has always been a challenge for deep learning models in the real world.[17] DRL, a branch of deep learning rapidly developed in the past 2 years and shows good performance in solving dynamic decision problems,[18–20] may have potential to mitigate this problem. DRL is often difficult to design and control; therefore, it only succeeded in a few cases, such as Atari Games[19] and Alpha Go,[18] which are usually repeatable games. Fortunately, there are basic principles (entering from the oesophagus into the gastrointestinal and then exit the oesophagus) during EGD procedures, so that we could use a lot of randomness to generate virtual EGD procedures and design a 'video game' to train and verify the DRL. In the present study, DCNN model achieved an accuracy of 88.70% in jam-free environment (still images). Using random forest filtering, a traditional reduction method, DCNN monitored blind spots with an accuracy of 82.61% in real EGD videos that contain lots of noises. After combining DRL with DCNN, the system achieved a much higher accuracy of 90.02% in videos. This is the first study using DRL in making medical decisions in human body environment, and this attempt

Table 3 The per cent of patients being ignored in each site between WISENSE and control group

Ignored sites	Number (%) WISENSE (n=153)	Control (n=150)	% Risk difference (95% CI)	P value
Oesophagus	0 (0.00)	0 (0.00)	NA	NA
Squamocolumnar junction	0 (0.00)	2 (1.33)	−1.33% (−4.74 to 1.14)	0.24
Antrum (G)	0 (0.00)	5 (3.33)	−3.33% (−7.57 to −0.84)	0.03
Antrum (P)	4 (2.61)	15 (10.00)	−7.39% (−13.56 to −2.10)	0.008
Antrum (A)	4 (2.61)	10 (6.67)	−4.05% (−9.56 to 0.77)	0.10
Antrum (L)	6 (3.92)	14 (9.33)	−5.41% (−11.63 to 0.20)	0.06
Duodenal bulb	1 (0.65)	6 (4.00)	−3.35% (−7.89 to 0.05)	0.06
Duodenal descending	0 (0.00)	9 (6.00)	−6.00% (−11.01 to −3.19)	0.002
Lower body (G)	4 (2.61)	26 (17.33)	−14.72% (−21.89 to −8.48)	<0.001
Lower body (P)	20 (13.07)	44 (29.33)	−16.26% (−25.36 to −7.17)	<0.001
Lower body (A)	11 (7.19)	28 (18.67)	−11.48% (−18.29 to −4.06)	0.003
Lower body (L)	8 (5.23)	45 (30.00)	−24.77% (−33.13 to −16.76)	<0.001
Middle-upper body (F, G)	4 (2.61)	8 (5.33)	−2.72% (−7.90 to 1.91)	0.244
Middle-upper body (F, P)	20 (13.07)	52 (34.67)	−21.59% (−30.87 to −12.20)	<0.001
Middle-upper body (F, A)	20 (13.07)	64 (42.67)	−29.59% (−38.97 to −19.97)	<0.001
Middle-upper body (F, L)	13 (8.50)	84 (56.00)	−47.50% (−56.20 to −38.06)	<0.001
Fundus (G)	4 (2.61)	13 (8.67)	−6.05% (−11.98 to −0.95)	0.02
Fundus (P)	13 (8.50)	32 (21.33)	−12.84% (−21.01 to −4.95)	0.002
Fundus (A)	22 (14.38)	26 (17.33)	−2.95% (−11.35 to 5.36)	0.49
Fundus (L)	29 (18.95)	61 (40.67)	−21.71% (−31.57 to −11.54)	<0.001
Middle-upper body (R, P)	10 (6.54)	26 (17.33)	−10.80% (−18.41 to −3.63)	0.003
Middle-upper body (R, A)	30 (19.61)	61 (40.67)	−21.06% (−30.97 to −10.84)	<0.001
Middle-upper body (R, L)	21 (13.73)	36 (24.00)	−10.27% (−19.14 to −1.48)	0.03
Angulus (P)	42 (27.45)	96 (64.00)	−36.55% (−46.50 to −25.70)	<0.001
Angulus (A)	19 (12.42)	80 (53.33)	−40.92% (−50.09 to −31.04)	<0.001
Angulus (L)	5 (3.27)	29 (19.33)	−16.07 (−23.51 to −9.42)	<0.001

Gut: first published as 10.1136/gutjnl-2018-317366 on 11 March 2019. Downloaded from http://gut.bmj.com/ on 13 March 2019 by guest. Protected by copyright.

successfully solved the problem of computer from perception to decision control in predicting gastric sites.

There are some limitations in the present study. First, while our results were obtained using Olympus and Fujifilm endoscopes, which have a 70%[31] and 14%[32] endoscope market share, respectively, we expect that the WISENSE will also be applied in endoscopes from other vendors. Theoretically, using the method of transfer learning, it could be achieved with little additional tuning of the algorithm.[33] It is possible for WISENSE to become a more universal EGD assistance system. Second, it is recommended that the EGD should last for at least 7 min on a patient without a previous gastroscopy for the last 3 years[1]. However, in the present study, the total time per procedure (even those cases that all 26 sites were observed) was less than the recommendations. It is an unusual phenomenon and we suppose that WISENSE, just like Ariadne's thread, enables endoscopists to conduct the procedure in a standardised way, and to easily realise which parts have not been observed yet, and these may guide them to accomplish the examination more quickly. However, further research should be conducted to verify this suppose.

In summary, WISENSE, a quality improving system for endoscopy based on deep learning achieves real-time monitoring on blind spots, timing and obtaining photodocumentation during EGD. WISENSE greatly reduced blind spot rate, increased inspection time and improved the completeness of photodocumentation in a randomised controlled trial. It could be a powerful assistant tool for mitigating skill variations among endoscopists and improving the quality of everyday endoscopy.

Author affiliations
[1]Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan, China
[2]Key Laboratory of Hubei Province for Digestive System Disease, Renmin Hospital of Wuhan University, Wuhan, China
[3]Hubei Provincial Clinical Research Center for Digestive Disease Minimally Invasive Incision, Wuhan University Renmin Hospital, Wuhan, China
[4]School of Resources and Environmental Sciences of Wuhan University, Wuhan, China

Contributors HGY conceived and supervised the overall study. LLW designed and conducted the experiments. JZ, WZ, PA, LS and XJ labelled image or video data. JL supervised the experiments. XH, GM, XinW, XL, JG and NC performed endoscopy in the clinical trial. SH, YC, XH and JL developed the system. DC, DG, XH and QD analysed the data. XZ, SL, XiaW, LX, XuW, JZ and MZ were involved in data collection. LLW wrote the manuscript. HGY is responsible for the overall content as guarantor. All authors approved the final version of the manuscript.

Funding This work was partly supported by the grant from the Research Funds for Key Laboratory of Hubei Province (No 2016CFA066), the National Natural Science Foundation of China (grant nos 81672387 [to HGY]) and the China Youth Development Foundation (grant no 81703030 [to QD]).

Competing interests None declared.

Patient consent for publication Obtained.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

REFERENCES

1 Bisschops R, Areia M, Coron E, *et al*. Performance measures for upper gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative. *Endoscopy* 2016;48:843–64.
2 Rutter MD, Rees CJ. Quality in gastrointestinal endoscopy. *Endoscopy* 2014;46:526–8.
3 Gado AS, Ebeid BA, Axon AT. Quality assurance in gastrointestinal endoscopy: An Egyptian experience. *Arab J Gastroenterol* 2016;17:153–8.
4 Zhang JG, Liu HF. Functional imaging and endoscopy. *World J Gastroenterol* 2011;17:4277–82.
5 Di L, Wu H, Zhu R, *et al*. Multi-disciplinary team for early gastric cancer diagnosis improves the detection rate of early gastric cancer. *BMC Gastroenterol* 2017;17:147.
6 Faigel DO. Quality, competency and endosonography. *Endoscopy* 2006;38(Suppl 1):65–9.
7 Malheiro R, de Monteiro-Soares M, Hassan C, *et al*. Methodological quality of guidelines in gastroenterology. *Endoscopy* 2014;46:513–25.
8 Faigel DO, Pike IM, Baron TH, *et al*. Quality indicators for gastrointestinal endoscopic procedures: an introduction. *Am J Gastroenterol* 2006;101:866–72.
9 Park WG, Cohen J. Quality measurement and improvement in upper endoscopy. *Tech Gastrointest Endosc* 2012;14:13–20.
10 Cohen J, Safdi MA, Deal SE, *et al*. Quality indicators for esophagogastroduodenoscopy. *Gastrointest Endosc* 2006;63(4 Suppl):S10–15.
11 Yao K. The endoscopic diagnosis of early gastric cancer. *Ann Gastroenterol* 2013;26:11.
12 Bretthauer M, Aabakken L, Dekker E, *et al*. Requirements and standards facilitating quality improvement for reporting systems in gastrointestinal endoscopy: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. *Endoscopy* 2016;48:291–4.
13 Torkamani A, Andersen KG, Steinhubl SR, *et al*. High-definition medicine. *Cell* 2017;170:828–43.
14 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
15 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.
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.
17 Tobin J, Fong R, Ray A, *et al*. Domain randomization for transferring deep neural networks from simulation to the real world. *IEEE* 2017:23–30.
18 Silver D, Huang A, Maddison CJ, *et al*. Mastering the game of Go with deep neural networks and tree search. *Nature* 2016;529:484–9.
19 Mnih V, Kavukcuoglu K, Silver D, *et al*. Human-level control through deep reinforcement learning. *Nature* 2015;518:529–33.
20 Mnih V, Kavukcuoglu K, Silver D, *et al*. Playing atari with deep reinforcement learning. *arXiv preprint arXiv* 2013;1312:5602.
21 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv* 2014;1409:1556.
22 Wen Z, Li B, Ramamohanarao K, *et al*. Improving Efficiency of SVM k-Fold Cross-Validation by Alpha Seeding. *AAAI* 2017:2768–74.
23 Glimcher PW. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc Natl Acad Sci U S A* 2011;108(Suppl 3):15647–54.
24 Fang M, Li Y, Cohn T. Learning how to active learn. *arXiv preprint arXiv* 2017;1708:02383.
25 Li J, Chai T, Lewis FL, *et al*. Off-policy interleaved Q-learning: optimal control for affine nonlinear discrete-time systems. *IEEE Trans Neural Netw Learn Syst* 2018:1–13.
26 Leung WK, Wu MS, Kakugawa Y, *et al*. Screening for gastric cancer in Asia: current evidence and practice. *Lancet Oncol* 2008;9:279–87.
27 Van Cutsem E, Sagaert X, Topal B, *et al*. Gastric cancer. *The Lancet* 2016;388:2654–64.
28 Miot HA. Sample size in clinical and experimental trials. *J Vasc Bras* 2011;10:275–8.
29 Huang Q, Shi J, Sun Q, *et al*. Clinicopathological characterisation of small (2cm or less) proximal and distal gastric carcinomas in a Chinese population. *Pathology* 2015;47:526–32.
30 Teh JL, Tan JR, Lau LJF, *et al*. Longer examination time improves detection of gastric cancer during diagnostic upper gastrointestinal endoscopy. *Clin Gastroenterol Hepatol* 2015;13:480–7.
31 Olympus Global. Olympus annual report 3. 2018 https://www.olympusglobal.com/ir/data/annualreport/pdf/ar2017e_A3.pdf.
32 Fujifilm Holding Corporation. Fujifilm holding corporation annual report. 2016 https://www.fujifilmholdings.com/en/investors/annual_reports/2016/pack/pdf/Annual-Report-2016.pdf.
33 Chang P, Grinband J, Weinberg BD, *et al*. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *AJNR Am J Neuroradiol* 2018;39:1201–7.
34 Urban G, Tripathi P, Alkayali T, *et al*. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology* 2018;155:1069–78.

Gut: first published as 10.1136/gutjnl-2018-317366 on 11 March 2019. Downloaded from http://gut.bmj.com/ on 13 March 2019 by guest. Protected by copyright.