A deep learning and natural language processing-based system for automatic identification and surveillance of high-risk patients undergoing upper endoscopy: A multicenter study

Jia Li,a,b,c, Shan Hu,d, Conghui Shi,a,b,c, Zehua Dong,a,b,c, Jie Pan, Yuwei Ai, Jun Liu,a,b,c, Wei Zhou,a,b, Yunchao Deng, Yanxia Li,a,b,c, Jingping Yuan, Zhi Zeng, Shuang Wen, a,b,c and Honggang Yu,a,b,c*

aDepartment of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan, Hubei 430060, PR China
bHubei Key Laboratory of Digestive System, Renmin Hospital of Wuhan University, Wuhan, Hubei 430060, PR China
cHubei Provincial Clinical Research Center for Digestive Disease Minimally Invasive Incision, Renmin Hospital of Wuhan University, Wuhan, Hubei 430060, PR China
dNational Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, Wuhan, Hubei 430079, PR China
eDepartment of Gastroenterology, Wenzhou Central Hospital, Wenzhou, Zhejiang 325000, PR China
fDepartment of Gastroenterology, The People’s Hospital of China Three Gorges University, The First People’s Hospital of Yichang, Yichang, Hubei 443000, PR China
gNursing Department of Renmin Hospital of Wuhan University, Wuhan, Hubei 430060, PR China
hDepartment of Pathology, Renmin Hospital of Wuhan University, Wuhan, Hubei 430060, PR China

Summary

Background Timely identification and regular surveillance of patients at high risk are crucial for early diagnosis of upper gastrointestinal cancer. However, traditional manual surveillance method is time-consuming, and current surveillance rate is below 50%. Here, we aimed to develop a surveillance system named ENDOANGEL-AS (automatic surveillance) for automatic identification and surveillance of high-risk patients.

Methods 7874 patients from Renmin Hospital of Wuhan University between May 1 and July 31, 2021 were used as the training set, 6762 patients between August 1 and October 31, 2021 as the internal test set, and 7570 patients from two other hospitals between August 1 and October 31, 2021 as the external test sets. We first extracted descriptions of abnormalities from endoscopic and pathological reports based on natural language processing techniques to identify individuals. Then patients were classified at nine risk levels according to endoscopic and pathological findings, and a deep learning model was trained to identify demarcation line (DL) in gastric low-grade intraepithelial neoplasia (LGIN) using 1561 white-light still images for risk stratification of gastric LGIN. Finally, patients undergoing upper endoscopy were classified and assigned one of ten surveillance intervals according to guidelines. The performance of ENDOANGEL-AS was evaluated and compared with physicians.

Findings Patient identification module achieved an accuracy of 100% and 99.91% in internal and external test sets, respectively. Risk level classification module achieved an accuracy of 100% and 99.85% in the internal and external test sets, respectively. DL identification module achieved an accuracy of 87.88%. ENDOANGEL-AS on surveillance interval assignment achieved an accuracy of 99.23% and 99.67% in internal and external test sets, respectively. ENDOANGEL-AS had significantly higher accuracy compared with physicians (99.00% vs 38.87%, p < 0.001). The accuracy (63.67%, p < 0.001) of endoscopists with the assistance of ENDOANGEL-AS was significantly improved.

Interpretation We established a surveillance system that can automatically identify patients and assign surveillance intervals with high accuracy and good transferability.
Funding This work was partly supported by a grant from the Hubei Province Major Science and Technology Innovation Project (2018-916-000-008) and the Fundamental Research Funds for the Central Universities (2042021kf0084).

Copyright © 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Upper gastrointestinal cancer; Surveillance; Endoscopy; Natural language processing; Deep learning

Introduction

Upper gastrointestinal (GI) cancers are one of the most common malignant tumors worldwide and remain a major health threat.1 The cascade of premalignant lesions preceding the development of gastric cancer (GC) includes atrophy, intestinal metaplasia (IM), and dysplasia.2,3 Barrett's esophagus and dysplasia are premalignant lesions of esophageal cancer (EC).4 Guidelines recommend that patients with high-risk lesions should undergo regular and timely surveillance, which can greatly improve early diagnosis and treatment rates.5–8 However, over 50% of high-risk patients do not receive endoscopy surveillance within recommended surveillance interval,9–11 due to a variety of factors, including lack of awareness of upper GI cancers, delay in informing patients of surveillance, and failure of physicians to follow the guidelines for surveillance recommendations. O'Connor et al. retrospectively reviewed 1055 patients with IM and found that only 48.1% of the patients had a further upper GI endoscopy, of which only 78 patients had surveillance endoscopies based on a prior finding of IM.10 Even among patients with low-grade intraepithelial neoplasia (LGIN) at higher risk, the surveillance rate was only about 40%.11 Accurate identification of patients who need surveillance and providing timely reminders of surveillance could potentially increase the surveillance rate of these patients.

With the rapid development of artificial intelligence (AI) in recent years, it has shown great value in gastrointestinal (GI) endoscopy. However, previous researches are mainly focused on the detection and diagnosis of abnormalities,12,13 rarely studies pay attention to the automatic surveillance system for high-risk patients with upper GI precancerous conditions. The traditional surveillance of patients mainly depends on manual reminders of clinicians, which leads to inconsistent and potentially inadequate management of high-risk patients due to inconsistent physician practices based on different guidelines.14,15 The combination of surveillance work with AI, which is good at handling mechanical and repetitive tasks, may address these surveillance-related issues.16

In this study, we developed an automatic surveillance system combining deep learning and natural language processing-based automatic surveillance system named ENDOANGEL-AS (automatic surveillance) for high-risk patients undergoing upper endoscopy. ENDOANGEL-AS performing significantly better than physicians can accurately identify high-risk patients and assign surveillance intervals according to text reports and images. To our best knowledge, this is the first study to develop an automatic surveillance system with high accuracy and good transferability in identifying high-risk patients undergoing upper endoscopy and assigning surveillance intervals.

Research in context

Evidence before this study
We searched PubMed for papers published from January 1, 2000 to April 1, 2022, with the keywords "artificial intelligence" OR "deep learning" AND "esophageal" OR "gastric" AND "endoscopy" AND "surveillance interval". Although timely and regular surveillance of patients at high-risk are crucial for early diagnosis of upper gastrointestinal cancer, the current surveillance rate is not satisfactory due to various reasons. While many studies confirm that artificial intelligence (AI) shows great value in gastrointestinal (GI) endoscopy, rarely study pays attention to the automatic surveillance system for high-risk patients with upper GI precancerous conditions.

Added value of this study
In the study, we developed a deep learning and natural language processing-based automatic surveillance system named ENDOANGEL-AS (automatic surveillance) for high-risk patients undergoing upper endoscopy. ENDOANGEL-AS performing significantly better than physicians can accurately identify high-risk patients and assign surveillance intervals according to text reports and images. To our best knowledge, this is the first study to develop an automatic surveillance system with high accuracy and good transferability in identifying high-risk patients undergoing upper endoscopy and assigning surveillance intervals.

Implications of all the available evidence
AI systems such as ENDOANGEL-AS have great potential in increasing the surveillance rate, improving the early diagnosis rate of upper GI cancers, and relieving the workload of physicians. Our system was able to automatically identify and assign surveillance intervals for high-risk patients undergoing upper endoscopy. It is expected to be used as a real-time surveillance reminder to promote timely and proactive surveillance of high-risk patients.
processing (NLP) techniques named ENDOANGEL-AS (automatic surveillance) for high-risk patients with upper GI precancerous conditions, which can automatically identify patients, classify patients into ten risk levels according to guidelines, and assigned corresponding standardized surveillance intervals from endoscopical and pathological reports. The performance of ENDOANGEL-AS was fully evaluated in internal and external tests and further compared with that of physicians in various departments. This study is expected to automatically remind patients and physicians of timely endoscopy surveillance to increase the surveillance rate, improve the early diagnosis rate of upper GI cancers, and relieve the workload of physicians.

Methods

Datasets

We retrospectively collected text endoscopic and pathological reports of 14,636 consecutive patients receiving endoscopy in Renmin Hospital of Wuhan University (RHWU, Medcare system [Qingdao, China]) from May 1 to October 31, 2021, of which 7874 patients from May 1 to July 31 for the training set and 6762 patients from August 1 to October 31 for the internal test set. We obtained endoscopic and pathological reports of 7570 consecutive patients for external test sets from Wenzhou Central Hospital (WCH, external test set 1, 5169 patients, Kayisoft system [Zhejiang, China]) and The First People’s Hospital of Yichang (FPHY, external test set 2, 2401 patients, DHC software system [Beijing, China]) during August 1 to October 31, 2021. The three electronic health record (EHR) systems used for each of the three sites were all semi-structured reporting systems. WCH took the Sydney criteria for gastric biopsies, and RHWU and FPHY took the targeted biopsy from endoscopically abnormal areas. The process for collecting datasets is summarized in Fig. 1. The representative raw data of endoscopy and pathology reports from each hospital is shown in Fig. S1. The characteristics of patients in test sets are demonstrated in Table S1.

Surveillance interval determination

To provide reasonable recommendations, two expert endoscopists (expert A and expert B, with more than 10 years of experience of endoscopy) and two pathologists (pathologist A and pathologist B) jointly classified surveillance intervals of the upper GI lesions into ten categories based on risk levels.

Fig. 1: The workflow of developing ENDOANGEL-AS. AS: automatic surveillance, RHWU: Renmin Hospital of Wuhan University, WCH: Wenzhou Central Hospital, FPHY: The First People’s Hospital of Yichang, ESD: endoscopic submucosal dissection, LGIN: low-grade intraepithelial neoplasia, DL: demarcation line, HGIN: high-grade intraepithelial neoplasia. * Wu L, Xu M, Jiang X et al. Real-time artificial intelligence for detecting focal lesions and diagnosing neoplasms of the stomach by white-light endoscopy (with videos). Gastrointest Endosc. 2022; 95(2):269-280.e6.
according to the guidelines issued by European Society for Gastrointestinal Endoscopy (ESGE) and Chinese Medical Association.1–4 (Fig. 1) If there was disagreement between the experts, a reassessment was carried out to reach a consensus as the ground truth.

The surveillance interval recommendations were as follows: 1) gastric high-grade intraepithelial neoplasia (HGIN) or cancer: immediate treatment, 2) gastric LGIN with a clear demarcation line (DL): 6 months, 3) gastric LGIN without a clear DL: 1 year, 4) gastric IM: 2–3 years, 5) severe atrophic gastritis: 1–2 years, 6) mild-to-moderate atrophic gastritis: 3 years, 7) esophageal HGIN or cancer: immediate treatment, 8) esophageal LGIN: 1–3 years, 9) Barrett’s esophagus: 3–5 years, 10) normal or other diseases: needless for surveillance or others. If the patient has multiple high-risk lesions, the highest risk level and the shortest surveillance interval shall be taken as the final recommended result.

Development of ENDOANGEL-AS
ENDOANGEL-AS consists of four modules: automatic patient identification module based on model 1 according to exclusion criteria, risk level classification module based on model 2, DL identification in gastric LGIN module based on model 3, and surveillance interval assignment module. The workflow and framework of developing ENDOANGEL-AS are shown in Figs. 1 and 2.

The construction framework of model 1 and model 2 was similar and illustrated in Fig. S2. The same two expert endoscopists (expert A and expert B) and two pathologists (pathologist A and pathologist B): The pathologists were invited to train pathology-related knowledge of upper gastrointestinal tumors and precancerous conditions before the expert endoscopists began annotating, so that the definition of upper gastrointestinal tumors and precancerous conditions could be unified when experts annotate pathological reports) annotated endoscopic and pathological reports as the ground truth to train models, and the annotations include entity and relationship annotations. If there was disagreement between the experts, a reassessment was carried out to reach a consensus as the ground truth. For model 1, the entity annotations were based on the exclusion criteria (described in the following sections), while for model 2, entity annotations were based on the risk level of lesions.

The entity annotations of the two models were to extract the keywords describing sites, degree of lesions, lesions, etc. (e.g., esophagus, mild, atrophic gastritis). Different categories of keywords were classified and annotated, for example, the keyword ‘esophagus’, ‘gastric’, ‘gastric body’, and ‘antrum’ all belong to the ‘sites’ category. The relationship annotations were rule-based (nearest location match) and associated different categories of keywords. Therefore, the keyword library and relationship dictionary related to the exclusion criteria and risk levels were obtained. The keywords were extracted from the reports and input into the keywords classifier that identified which category the keywords belong to, and the rule-based relationship matcher associates different categories of keywords. Finally, the rule-based patient classifier was constructed by combining the classification rules given by experts with the relationship dictionary to identify whether the exclusion criteria were met or which risk level belongs. Table S2 describes the entities and their descriptions in reports.

Development of automatic patient identification module based on model 1
Model 1 automatically identified patients according to descriptions of exclusion criteria from raw endoscopic and pathological reports of the training set based on NLP. The exclusion criteria were as follows: 1) without pathological results in the upper GI tract, 2) therapeutic endoscopy or with a history of previous gastrectomy, esophagectomy, or endoscopic submucosal dissection (ESD), 3) the degree of dysplasia was not noted in the pathology reports, 4) the specific biopsy sites of atrophic gastritis involving body were not presented in the endoscopic or pathological reports. Multiple exclusion criteria may be met in reports of one patient, which was handled by limiting the model to process each exclusion criterion in sequence.

Development of risk level classification module based on model 2
Model 2 classified patients correctly included in model 1 into nine risk levels based on endoscopic and pathological reports. The classification of all patients was mainly based on pathological reports, while the endoscopic reports were only to assist in determining the biopsy site. Since the DL of the gastric LGIN patients cannot be identified from reports, they were temporarily classified into one level in this model.

A report may describe multiple lesions at different anatomical sites, in which case each lesion can be processed in sequence, and they might be distinguished by lesion descriptions and anatomical sites. A new lesion discovery can be created each time a new anatomic site, degree, or histology is encountered. The highest risk level will be taken as the basis for the final risk level classification.

Development of DL identification in gastric LGIN module based on model 3
To identify the DL in gastric LGIN patients, we constructed model 3 based on supervised learning (SL) and semi-supervised learning (SSL) respectively. ResNet-50 was used for the development of the SSL model and the SL model using 1561 white-light images from 387 LGIN patients histologically confirmed retrospectively.
collected between Jan 3, 2017 and Jul 3, 2021 in RHWU. Before training the DL identification model, the lesions in the whole images were labeled with rectangular boxes by YOLO for object detection assessing the presence of gastric lesions and localizing them of our previous study. Then the rectangular boxes were enlarged by...
1.2 times to better display the DL of lesions, and the lesions were cropped out according to the enlarged boxes. The same two expert endoscopists (expert A and expert B) reviewed the cropped LGIN images and evaluated the DL of each LGIN lesion (when multiple LGIN lesions appeared in an image, multiple rectangular boxes were used). If there was disagreement between the two endoscopists, a reassessment was carried out to reach a consensus as the ground truth to train model 3.

Finally, 1697 cropped LGIN images were obtained. Then the cropped LGIN images were randomly assigned to a training set (1466 cropped images from 345 LGIN patients) and a test set (231 cropped images from 42 LGIN patients) by the patient with a ratio of about 7:1. The workflow of developing model 3 is illustrated in Fig. 1. The framework of developing model 3 is shown in Fig. S3. The performance of the SL model and SSL model trained with labeled training sets in increments of 10% was evaluated in the still image test set. Finally, an optimization model with the best Youden index was selected for model 3.

The SSL model was constructed based on the Mean Teacher architecture within PyTorch 1.4.0 and SGD optimizer with an initial learning rate of 0.001. The SL model was constructed using the Keras library (v2.1.5) with Tensorflow 1.3.0 backend and Adam optimizer with an initial learning rate of 0.001. For each epoch, every 16 images were divided as a batch to be placed into the SSL model and every 64 images were divided as a batch to be placed into the SL model.

Development of surveillance interval assignment module
The final module combined the above three models to classify the patients into one of ten surveillance intervals (Figs. 1 and 2).

Model 1 identified patients in raw data according to the exclusion criteria, and model 2 classified risk levels of patients according to determined surveillance interval categories based on NLP. If it was a gastric LGIN patient, the images of LGIN lesions were selected according to the sites of gastric LGIN described in pathological reports. Then YOLO labeled lesions with boxes in the selected images. After the boxes were enlarged by 1.2 times, model 3 identified the DL of the labeled lesions in the boxes. Finally, the surveillance interval assignment module recommended appropriate surveillance intervals.

Video 1 showed surveillance procedures for patients in clinical practice. ENDOANGEL-AS was developed using Python (Python Software Foundation, Wilmington, Del, USA) and the spaCy framework (Explosion, Berlin, Germany), an open-source library for NLP.

Test of ENDOANGEL-AS
We first tested the performance of ENDOANGEL-AS including four modules using an internal test set from RHWU. Then we assessed the robustness of the system using external test set 1 and external test set 2. All the endoscopic and pathological reports were jointly evaluated by the same two expert endoscopists (expert A and expert B) and the same two pathologists (pathologist A and pathologist B) as the gold standard.

Comparing the performance of ENDOANGEL-AS and physicians
To compare the performance between ENDOANGEL-AS and physicians, we further selected inpatients from the internal test set who did not meet the above exclusion criteria. These inpatients had traceable complete medical records and surveillance interval recommendations given by physicians, which were obtained by reviewing the advice after discharge in the EHR only available to hospitalized patients. A subgroup analysis was also performed to compare the performance of physicians in different departments with that of ENDOANGEL-AS. The specialties of the physicians are shown in Table S3. To explore the assistance ability of ENDOANGEL-AS, we trained three endoscopists, who were not enrolled in the annotation of the reports or images, on surveillance guidelines of the upper GI lesions, and tested the performance of endoscopists’ independent surveillance recommendations of 100 randomly selected patients after at least 2 weeks. The three endoscopists were asked to make a surveillance recommendation again on the same test set with the assistance of ENDOANGEL-AS after a 2-week washout period, and compared it with their independent performance.

Ethics statement
This study was approved by the Ethics Committee of RHWU, WCH, and FPHY. For retrospective datasets, informed consent was exempted by the institutional review boards.

Statistical analysis
The performance of ENDOANGEL-AS was evaluated by metrics including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the curve (AUC). The optimal threshold of the receiver operating characteristic (ROC) curve was determined by Youden index. The performance of physicians was evaluated by accuracy and McNemar test was used to compare the performance of ENDOANGEL-AS and physicians. When the output of the ENDOANGEL-AS was identical to the ground truth given by the experts, it was considered correct. And when the output was different from the ground truth, it was considered a mistake. P-values <0.05 were considered statistically significant.
Role of the funding source
The funder had no role in study design, data collection, data analysis, data interpretation, writing of the report, or decision to submit the paper for publication. All authors had access to the raw datasets and accept responsibility for the decision to submit for publication.

Results
The performance of model 1 on identifying patients from raw data
Model 1 on identifying patients achieved an overall accuracy of 100% (6762/6762, 95% confidence interval [CI]: [99.94–100.00%]) on the internal test set. In external test set 1, the overall accuracy was 99.92% (5165/5169, 95% CI: [99.80–99.97%]). Four patients were mistakenly excluded due to: 1) unidentified duodenal lesions from pathological reports (n = 3); 2) mistakenly identified esophageal HGIN as dysplasia without degree description, because the pathological report described “(Esophagus) Squamous dysplasia, and can’t rule out HGIN”, the experts rigidly classified the patient as esophageal HGIN, while ENDOANGEL-AS did not understand “can’t rule out HGIN” and classified the description of “squamous dysplasia” as “dysplasia without degree description” (n = 1). The overall accuracy on external test set 2 was 99.88% (2398/2401, 95% CI: [99.64–99.96%]) and three patients were mistakenly excluded due to: 1) failed to identify gastritis or duodenogastric metaplasia (n = 2); 2) mistakenly identified gastric hemostasis as ESD (n = 1). The detailed results of the identification are shown in Table 1.

The performance of model 2 on classifying the risk levels from text reports
The overall accuracy of model 2 on risk level classification from text reports for patients was 100% (780/780, 95% CI: [99.51–100.00%]) in the internal test set, 99.92% (4771/4775, 95% CI: [99.79–99.97%]) in the external test set 1, and 99.05% (415/419, 95% CI: [97.58–99.63%]) in the external test set 2. A total of eight patients were misclassified due to: 1) unidentifed atrophic gastritis (n = 1); 2) mistakenly identified atrophic gastritis of angulus as atrophic gastritis of body (n = 2); 3) mistakenly identified duodenal dysplasia as gastric dysplasia (n = 3); 4) mistakenly identified esophageal squamous papilloma as EC (n = 1); 5) The coexistence of Barrett’s esophagus and gastric IM only identified Barrett’s esophagus (n = 1). The classification accuracy of each risk level is shown in Table 2. To reduce the potential impact of quantitative imbalance in the risk levels, a subgroup analysis was performed after excluding Category 10. The results show that the overall accuracy of model 2 in the internal test set, external test set 1, and external test set 2 were 100% (324/324, 324/324, 324/324), respectively.

| Datasets | Overall (n/N, 95% CI) | No upper gastrointestinal pathology (n/N, 95% CI) | No dysplasia degrees (n/N, 95% CI) | No biopsy sites (n/N, 95% CI) |
|----------|------------------------|---------------------------------------------|---------------------------------|-------------------------------|
| Internal test set (n = 6762) | 100% (6762/6762, 99.94–100.00%) | 100% (5432/5432, 99.93–100.00%) | 100% (544/544, 99.93–100.00%) | 100% (4/4, 99.94–100.00%) |
| External test set 1 (n = 5169) | 99.92% (5165/5169, 99.92–100.00%) | 100% (185/185, 99.98–100.00%) | 100% (5/5, 99.94–100.00%) | 100% (4/4, 99.94–100.00%) |
| External test set 2 (n = 2401) | 99.30% (2398/2401, 99.30–100.00%) | 100% (1781/1781, 99.90–100.00%) | 100% (191/191, 99.61–100.00%) | 100% (4/4, 99.94–100.00%) |

Table 1: The accuracy of model 1 on identifying patients from raw data.
### Table 2: The accuracy of ENDOANGEL-AS on risk level classification and surveillance interval assignment, the accuracy of ENDOANGEL-AS and physicians on surveillance interval assignment of patients, and the performance of endoscopists on surveillance recommendations.

| Datasets | Overall (95% CI) | Category 1 (95% CI) | Category 2 (95% CI) | Category 3 (95% CI) | Category 4 (95% CI) | Category 5 (95% CI) | Category 6 (95% CI) | Category 7 (95% CI) | Category 8 (95% CI) | Category 9 (95% CI) | Category 10 (95% CI) |
|----------|-----------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| ENDOANGEL-AS on risk level classification based on reports | | | | | | | | | | | |
| Internal test set (n = 270) | 100% (780/780, 99.51-100.00%) | 100% (57/57, 93.69-100.00%) | 100% (38/38, 90.82-100.00%) | 100% (198/198, 98.10-100.00%) | 100% (4/4, 51.01-100.00%) | 100% (77/77, 64.57-100.00%) | 100% (15/15, 79.61-100.00%) | 100% (5/5, 56.55-100.00%) | - | 100% (456/456, 99.16-100.00%) | |
| External test set (n = 4775) | 100% (4771/4775, 99.79-100.00%) | 100% (65/65, 94.62-100.00%) | 100% (16/16, 80.64-100.00%) | 100% (12/12, 75.75-100.00%) | 100% (1/1, 97.74% | 100% (12/12, 77.19-100.00%) | 100% (14/14, 94.69-100.00%) | 100% (4/4, 51.01-100.00%) | 100% (4/4, 41.67-100.00%) | 99.97% (3525/3526, 99.84-99.99%) | |

**ENDOANGEL-AS on surveillance interval assignment based on both reports and images**

**ENDOANGEL-AS and physicians on surveillance interval assignment of patients**

All departments (n = 101)

Internal test set (n = 270)

- ENDOANGEL-AS: 99.00% (298/301, 97.11-99.66%)
- Physicians: 90.11% (300/334, 89.28-90.93%)
- Gastroenterology department (n = 174)
  - ENDOANGEL-AS: 94.43% (167/174, 93.34-95.52%)
  - Physicians: 87.76% (152/172, 86.65-88.88%)
- Other departments (n = 127)
  - ENDOANGEL-AS: 90.26% (113/127, 89.19-91.35%)
  - Physicians: 87.55% (110/127, 86.47-88.56%)
- The performance of endoscopists on surveillance recommendations:
  - Without ENDOANGEL-AS: 30.00% (93/310, 29.00-31.00%)
  - With ENDOANGEL-AS: 63.67% (191/300, 62.56-65.00%)

Risk level classification was based on reports, in which gastric LGIN with DL and gastric LGIN without DL were classified into one level because the DL cannot be identified from the images. Surveillance interval assignments were based on both reports and images, in which the DL can be identified from the images. Category 1 is gastric LGIN with clear DL, and immediate treatment is recommended. Category 2 is gastric LGIN with DL, and surveillance interval is 6 months. Category 3 is gastric LGIN without DL and the surveillance interval is 1 year. Category 4 is gastric IM and the surveillance interval is 1 to 2 years. Category 5 is severe atrophic gastritis and the surveillance interval is 1 to 2 years. Category 6 is mild-to-moderate atrophic gastritis and the surveillance interval is 1 to 2 years. Category 7 is epithelial LGIN and the surveillance interval is 6 months. Category 8 is epithelial LGIN and the surveillance interval is 1 to 2 years. Category 9 is Barrett’s esophagus and the surveillance interval is 3 to 5 years. Category 10 is normalities or other diseases and the surveillance interval is 6 months. Category 11 is automatic surveillance, C: confidence interval, HGN: high-grade intraepithelial neoplasia, LGIN: low-grade intraepithelial neoplasia, DL: demarcation line, IM: intestinal metaplasia. *Significant difference between the target groups and ENDOANGEL-AS (p < 0.05). The numerator is the correct total number of patients of three endoscopists, and the denominator is the total number of patients of three endoscopists. **Significant difference between the results of without ENDOANGEL-AS and with ENDOANGEL-AS (p < 0.05).
98.83–100.00%), 99.76% (1246/1249, 99.30–99.92%) and 99.03% (204/206, 96.53–99.73%) respectively (Table S4).

The performance of model 3 on identifying DL in gastric LGIN

The AUC of SL model was 87.83% (95% CI: [83.18–92.47%]), and it achieved an accuracy of 80.52% (186/231, 95% CI: [74.93–85.11%]), a sensitivity of 80.00% (92/115, 95% CI: [71.77–86.29%]), and a specificity of 81.03% (94/116, 95% CI: [72.95–87.12%]) on identify clear DL in gastric LGIN. The AUC of SSL model, which performed the best when trained with a 20% labeled training set, was 95.29% (95% CI: [92.70–97.88%]). SSL model achieved an accuracy of 95.26% (203/231, 95% CI: [91.94–97.94%]), a sensitivity of 85.22% (98/115, 95% CI: [77.60–92.84%]), a specificity of 90.52% (105/116, 95% CI: [83.82–94.62%]). The detail of the performance of model 3 is shown in Fig. 3 and Table S5.

The SSL model performing better was selected for model 3. The accuracy of model 3 was 84.21% (32/38, 95% CI: [68.58–92.56%]) in the internal test set, 87.50% (14/16, 95% CI: [63.98–96.50%]) in the external test set 1, and 100% (1/1, 20.65–100.00%) in the external test set 2 (Table 2). The main error reason for the DL identification model was that YOLO mistakenly labeled mucus or hemorrhage and lesions were not in the center of the visual field.

The performance of ENDOANGEL-AS on surveillance interval assignment

Combined with model 1 and model 2 based on reports and model 3 based on images, the overall accuracy of ENDOANGEL-AS on surveillance interval assignment was 99.23% (774/780, 98.33–99.65%) in the internal test set, 99.79% (4769/4779, 99.61–99.89%) in the external test set 1, and 98.34% (415/422, 96.61–99.19%) in the external test set 2. The accuracy of the surveillance interval assignment is shown in Table 2. When the DL of gastric LGIN was not considered, ENDOANGEL-AS classified patients into nine categories, and the overall accuracy of ENDOANGEL-AS in the internal test set, external test set 1, and external test set 2 were 100% (780/780, 99.51–100.00%), 99.83% (4771/4779, 99.67–99.91%) and 98.34% (415/422, 96.61–99.19%) respectively (Table S6). The overall accuracy of ENDOANGEL-AS in the internal test set, external test set 1, and external test set 2 after excluding Category 10 were 98.15% (318/324, 96.02–99.15%), 99.52% (1244/1250, 98.96–99.78%) and 99.03% (204/206, 96.53–99.73%) respectively (Table S4).

The processing computation time of ENDOANGEL-AS was 22.85 ± 39.12 ms in the testing sets. Notably, the processing computation time was longer in gastric LGIN patients (376.21 ± 273.82 ms) than in those without LGIN (20.12 ± 0.33 ms) because the system needs to invoke more models to label the lesions and identify the demarcation line of the lesions.
Comparing the performance of ENDOANGEL-AS and physicians

ENDOANGEL-AS showed significantly higher accuracy when compared with physicians both in gastroenterology and non-gastroenterology departments (overall: 99.00% vs 38.87%, p < 0.001, gastroenterology department: 99.43% vs 24.71%, p < 0.001, other departments: 98.43% vs 58.27%, p < 0.001) (Table 2). Compared with physicians in other departments, physicians in the gastroenterology department were more willing to give surveillance interval recommendations (85.50% [112/131] vs 39.62% [21/53], p < 0.001). The accuracy of endoscopists at different levels was significantly lower than that of ENDOANGEL-AS (Table S7). The most common reasons leading to the mistakes of physicians include that they gave surveillance interval recommendations to patients who did not need (46.74%, 86/184), did not give recommendations to patients who should be given (27.72%, 51/184), or shortened the surveillance intervals (23.91%, 44/184). Among them, all patients with gastric IM (0/56) were not given correct surveillance intervals by physicians, and a significant proportion of giving surveillance intervals to patients when not indicated. The reasons leading to the mistakes are shown in Table S8.

As shown in Table 2, the accuracy of the endoscopists with the assistance of ENAOANGEL-AS was significantly improved (63.67% vs 30.00%, p < 0.001). The lesion with the most obvious improvement in accuracy was gastric IM (94.87% vs 66.67%, p < 0.001). The details of the performance of the three endoscopists with the assistance of ENAOANGEL-AS are shown in Table S9.

Discussion

In the study, we developed a deep learning and NLP-based automatic surveillance system named ENDOANGEL-AS for high-risk patients undergoing upper endoscopy. ENDOANGEL-AS performing significantly better than physicians can accurately identify high-risk patients and assign surveillance intervals according to text reports and images. To our best knowledge, this is the first study to develop an automatic surveillance system with high accuracy and good transferability, which is expected to increase the surveillance rate, improve the early diagnosis rate of upper GI cancers, and relieve the workload of physicians.

Early identification of high-risk patients and generation of management strategies to slow or prevent upper GI cancer progression can reduce the incidence and mortality. In addition, early detection of upper GI cancer and endoscopic treatment are more cost-effective for patients than surgical resection. Although surveillance is vital and there are numerous guidelines and comments advocating regular and timely surveillance, the current rate of surveillance is not satisfactory. It is imperative to accurately identify high-risk patients and assign surveillance intervals.

ENDOANGEL-AS we proposed achieved impressive accuracy across three different EHR systems at three hospitals, which can automatically identify high-risk patients and assign surveillance intervals according to the guidelines from massive endoscopic and pathological reports. It was challenging to burden endoscopists with additional intensive surveillance work requiring a great effort in the clinic. Considering that endoscopists have heavy clinical tasks and a large number of endoscopes performed daily. And the identification and tracking of high-risk patients and the determination of surveillance intervals need a synthesis of multiple data which are usually available at different times and in various locations. ENDOANGEL-AS simplified the clinical process by automatically identifying high-risk patients and assigning surveillance intervals by identifying the description in the reports. The automatic management for high-risk patients has the advantages of high efficiency, short time consumption, and no fussiness. The remarkable performance of ENDOANGEL-AS at three hospitals implicated that it has considerable transferability. Based on semi-structured endoscopic and pathological reports, ENDOANGEL-AS can perform well even when applied to other systems. It had great potential in reducing the workload of endoscopists and improving the status of the low surveillance rate.

The surveillance interval determined by ENDOANGEL-AS based on guidelines was more objective. In clinical practice, physicians gave surveillance intervals according to endoscopic and pathological reports. Whereas there are many guidelines of surveillance interval recommendations for high-risk patients with upper GI precancerous conditions, and each endoscopist may have a different understanding of different guidelines. As a result, the consistency between endoscopists is poor in the assignment of surveillance intervals. ENDOANGEL-AS is based on recommendations for surveillance intervals given in the guidelines issued by ESGE and Chinese Medical Association. As an auxiliary surveillance system, ENDOANGEL-AS could not initially determine the biopsy site, but only provides physicians with surveillance interval advice based on the maximum available information and clinical guidelines. With the help of ENDOANGEL-AS, the disease progression due to incorrect surveillance interval assignment after pure referring to one guideline was reduced, and the confusion caused by different surveillance intervals recommended by multiple guidelines was avoided, which may improve the accuracy and consistency of endoscopists. Moreover, we evaluated the performance of the endoscopists with the assistance of ENAOANGEL-AS, and compared it with their independent performance. The comparison results showed that ENAOANGEL-AS could effectively improve...
the performance of endoscopists, which demonstrates the additional value of the ENAOANGEL-AS in auxiliary surveillance.

In the comparison of ENDOANGEL-AS and physicians, the proportion of physicians giving the correct surveillance intervals was unsatisfactory, especially for patients with gastric IM or without indication. It is known that the cascade of premalignant lesions prior to the development of gastric cancer includes atrophy, IM, and dysplasia. Therefore, in the consciousness of physicians, gastric IM is a more serious precancerous condition than atrophy, and they tend to recommend more aggressive surveillance intervals. However, the surveillance guidelines are evidence-based, and the surveillance interval of gastric IM (2–3 years) is longer than severe atrophic gastritis (1–2 years), which is different from the impression of some physicians and leads to a significant proportion of incorrect surveillance guidelines related to gastric IM. Furthermore, although guidelines recommend that patients without atrophy do not need surveillance, in clinical practice, physicians may prefer to have patient surveillance regularly, both to minimize the risk and increase patient visits. It actually results in the waste of medical resources, which is also one of the issues that ENDOANGEL-AS aims to solve effectively.

In addition, ENDOANGEL-AS can be used as a real-time surveillance reminder to promote timely and proactive surveillance of high-risk patients. After the patients undergo upper endoscopy, ENDOANGEL-AS can automatically obtain the examination results of the patients from EHR systems and enter the status of the surveillance recommendations given by the system, the patients enter the status of “Pending Surveillance”. ENDOANGEL-AS can automatically send a short message and make a phone call to inform patients before the patient needs to undergo surveillance endoscopy (the specific time of sending a reminder can be set freely), then the patients enter the status of “Informed”. When completing the surveillance endoscopy, the patients entered the status of “Completed Surveillance”. Many studies had concluded that enhanced surveillance by telephone and message can improve surveillance rate, adherence, and therapeutic efficiency, which required significant medical resources. ENDOANGEL-AS can automatically remind the patients and endoscopists at certain key time points before the surveillance, which was more acceptable and reduced overly frequent and unnecessary surveillance of patients in clinical practice.

In this study, the accuracy of identifying the description from reports based on NLP was slightly higher than that in previous studies. NLP is commonly used to generate structured information from unstructured free text for a variety of important clinical and research tasks. In the field of digestive endoscopy, NLP is mainly used to improve the quality of colonoscopy (accuracy was 0.89), predict colonoscopy surveillance intervals (accuracy was about 92%), and predict the risk of colorectal cancer (AUC was about 71.6%). Rare studies explored the potential value of NLP in upper GI endoscopy. We applied NLP for the first time in surveillance interval determination in high-risk patients with the upper GI precancerous conditions and achieved high accuracy (average accuracy was about 99%). ENDOANGEL-AS with slightly higher accuracy possibly because we mainly used semi-structured pathological reports with more standardized terminology as the basis for developing the system. The reports were structured based on NLP, which can quickly track and extract important information, helping endoscopists to notice useful information in the reports.

There are some important strengths to our study. Although AI is widely used in the field of GI endoscopy, deep learning has not been fully evaluated for the DL identification of gastric LGIN. Gastric LGIN with clear DL has a high risk of malignancy, and HGIN or EGC may already be present in the non-biopsy area. Therefore, gastric LGIN patients with clear DL need more frequent surveillance. To identify DL in gastric LGIN, we constructed the model based on SL and SSL, which achieved high accuracy. Moreover, we explore the performance of the SL model and SSL model, and find that the SSL model with only a 20% labeled training set performed better than the SL model with massive annotations. The results indicated that SSL may reduce the need for the amount of annotation data and achieve satisfactory performance compared to SL. In future work, the performance of SSL should be further explored to improve the efficiency of medical AI development.

There are several limitations to this study. First, although this study is a multi-center study, the selected centers are all in China. The surveillance interval recommendation was based on the guidelines issued by ESGE and Chinese Medical Association. It can be inferred that considerable performance can be achieved with our approach even if the guidelines are adjusted for the disease spectrum in different countries. At the same time, this should be taken into account when ENDOANGEL-AS is applied in other countries and populations. Second, this is a retrospective study to verify the accuracy of ENDOANGEL-AS in assigning surveillance intervals, and we should conduct a clinical trial to verify the effect of ENDOANGEL-AS on improving surveillance rate in the future to make ENDOANGEL-AS more convincing and reliable. Third, ENDOANGEL-AS relied on accurate pathological results from endoscopists and pathologists, and differences in biopsy protocols could potentially have an impact on the results of the endoscopy and pathology and further affect the surveillance interval for each patient. We couldn’t get pathological results from unbiopsied sites, which may lead to the underestimation of lesions. However, even in the context of extensive biopsies,
sampling errors and the possibility of higher-risk lesions at unbiopsied sites cannot be ruled out. In the future, the development of a standardized biopsy auxiliary system can be tried and combined with ENDANGEL-AS, which will effectively address the problems. Fourth, there was a quantitative imbalance of ENDANGEL-AS on risk level classification and surveillance interval assignment. Although we conducted a subgroup analysis after excluding Category 10 with a large majority of cases and obtained impressive performance, the number of each category was also not well balanced, which may have a potential impact on the accuracy of the ENDANGEL-AS. Although the dataset of this study was the real clinical cohort, large-scale and multi-center validation should be conducted in the future to make the results more convincing. Fifth, ENDANGEL-AS was mainly based on semi-structured reports. Although we tested the system in three hospitals with EHR systems and achieved promising performance, there may still be other hospitals using free-text reports. However, the guidelines indicate that endoscopy reporting systems should be integrated into hospital patient record systems, structured data is recommended and free text shall be restricted to a minimum. In an endoscopy report, the lesion area and features should be included. A pathologic examination should report histologic changes in each biopsy specimen in order to provide the clinicians with more detailed information. Therefore, standardized semi-structured reports are clinically required. Furthermore, the basement of ENDANGEL-AS on the quality of reports is conducive to the quality control of reports. Sixth, ENDANGEL-AS was developed and validated based on experts as the gold standard, and achieved satisfactory performance. However, it had to be admitted that there may be inter-observer and intra-observer variability in the assessment of experts, which may lead to the failure to correctly identify all high-risk patients. Therefore, potential implicit uncertainties should be taken into account when interpreting ENDANGEL-AS and its outcomes.

In conclusion, this study proposed an automatic surveillance system named ENDANGEL-AS aimed at accurately identifying high-risk patients undergoing upper endoscopy and assigning surveillance intervals. ENDANGEL-AS, which had better performance than physicians, had a promising role in automating surveillance reminders to patients and physicians and improving the surveillance rate.

Contributors
Honggang Yu and Lianlian Wu were responsible for conceiving and designing the study; Jia Li, Shan Hu, and Conghui Shi were responsible for training and testing the models; Jia Li, Shan Hu, Conghui Shi, Zehua Dong, Liu Jun, Yaowei Ai, Jie Pan, Wei Zhou, Yunchao Deng, Yanxia Li, Jingjing Yuan, and Zhi Zeng were responsible for collecting and reviewing data and images; Jia Li was responsible for collecting, collating and analyzing the data; Jia Li, Shan Hu, and Conghui Shi were responsible for writing the manuscript; Jia Li, Shan Hu, Conghui Shi, and Lianlian Wu were responsible for revising the manuscript; Honggang Yu was responsible for performing extensive editing of the manuscript; all authors reviewed and approved the final manuscript for submission. All authors were involved in data acquisition, general design of the trial, interpretation of the data, and critical revision of the manuscript. We ensured that all the authors had access to all the raw datasets. Jia Li, Shan Hu, Conghui Shi, Lianlian Wu, and Yu Honggang have verified the data. All authors contribute to the critical revision of the manuscript.

Data sharing statement
Individual de-identified participant data that underlie the results reported in this article and study protocol will be shared for investigators after article publication. To gain access, data requesters will need to contact the corresponding author.

Declaration of interests
We declare no competing interests.

Acknowledgments
This work was partly supported by a grant from the Hubei Province Major Science and Technology Innovation Project (2018-916-000-008) and the Fundamental Research Funds for the Central Universities (20140211k0084).

Appendix A. Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1016/j.clcm.2022.101704

References
1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
2. Pasechnikov V, Chukov S, Fedorov E, Kikuste I, Leja M. Gastric cancer: prevention, screening and early diagnosis. World J Gastroenterol. 2014;20(38):11384–11362.
3. Yoon H, Kim N. Diagnosis and management of high risk group for gastric cancer. Gut Liver. 2015;9(1):5–17.
4. Pennathur A, Gibson MK, Jobe BA, Luketich JD. Oesophageal carcinoma. Lancet. 2013;381(9864):400–412.
5. National Clinical Research Center for Digestive Disease (Shanghai), National Early Gastrointestinal Center Prevention & Treatment Center Alliance, Helicobacter Pylori Group, et al. Chinese consensus on management of gastric epithelial precancerous conditions and lesions (2020). Chin J Dig. 2020;40(11):731–741.
6. Pimentel-Nunes P, Libánio D, Marcos-Pinto R, et al. Management of epithelial precancerous conditions and lesions in the stomach (MAPS I): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. Endoscopy. 2019;51(4):365–388.
7. ASGE STANDARDS OF PRACTICE COMMITTEE, Qunsey A, B, Sultan S, et al. ASGE guideline on screening and surveillance of Barrett’s esophagus. Gastrointest Endosc. 2019;90(1):315–359.e2.
8. National Digestive Endoscopy Improvement System, National Clinical Research Center for Digestive Disease (Shanghai), National Early Gastrointestinal Center Prevention, et al. Chinese consensus on management of early esophageal cancer and precancerous lesions screening (2019, Xinxiang). Chin J Digest Endosc. 2019;36(11):793–801.
9. Chapelle N, Péron M, Quinéhervé L, et al. Long-term follow-up of gastric precancerous lesions in a low GC incidence area. Clin Transl Gastroenterol. 2020;11(12):e00217.
10. O’Connor A, Bowden A, Farrell E, et al. Risk of progression of gastric intestinal metaplasia is significantly greater when detected in both antrum and body. Dig Dis Sci. 2021;66(10):3470–3475.
11. Xiao S, Liu H, Xue Y, et al. Long-term outcome of gastric mild-moderate dysplasia: a real-world clinical experience. Clin Gastroenterol Hepatol. 2022;20(6):1259–1268.e7.

www.thelancet.com Vol 53 November, 2022
