

基于多标签分类的胃镜检查完整度评估 人工智能模型的构建与验证

林琳^{1,2} 姚双喆¹ 宋岩¹ 徐昕¹ 惠洋洋¹ 王玉峰³ 冯月³ 穆金宝³ 卢祎⁴
陈鑫¹ 王邦茂¹ 郑忠青¹

¹天津医科大学总医院消化内科,天津 300052;²天津医科大学总医院空港医院消化内科,天津 300308;³天津市御锦人工智能医疗科技有限公司,天津 300384;⁴中山大学附属第六医院消化内镜科,广州 510655

通信作者:王邦茂,Email:mwang02@tmu.edu.cn;郑忠青,Email:zzheng01@tmu.edu.cn

【摘要】 目的 评估利用 ResNeSt-50 网络构建的多标签分类模型(下文称为 Endosmart)在胃镜检查部位识别中的应用价值。方法 收集天津医科大学总医院空港医院及天津市南开医院 2018 年 1 月至 2022 年 7 月间的胃镜检查图片 10 172 张,包括 24 个胃部位及 13 617 个标签,其中 8 501 张用于深度学习构建胃镜部位识别多标签分类模型 Endosmart,实现输入的胃镜图片同时输出多个解剖部位的多标签预测;使用 1 671 张胃镜图片进行内部验证;收集天津医科大学总医院 2022 年 8 月至 2024 年 12 月间的 100 例胃镜检查视频片段作为外部验证;分别评估 Endosmart 在内部图片验证集及外部视频验证集中对胃各部位识别的精确度、召回率和 F1 分数;并在视频验证集中,对比该模型与 4 名中年资内镜医师的平均识别效能。结果 在内部图片验证集中,Endosmart 识别胃各部位的平均精确度为 94.2% (95%CI: 92.6%~95.5%),召回率为 90.3% (95%CI: 88.7%~91.7%),F1 分数为 86.2% (95%CI: 84.2%~88.2%);在外部视频验证集中,Endosmart 对胃各部位识别的精确度、召回率及宏平均 F1 分数分别为 95.5% (95%CI: 94.0%~96.9%)、93.7% (95%CI: 92.4%~95.0%) 和 94.5% (95%CI: 93.3%~95.6%);4 名中年资内镜医师对胃各部位整体识别的平均精确度为 95.5% (95%CI: 93.9%~96.9%),召回率为 90.7% (95%CI: 89.1%~91.8%),宏平均 F1 分数为 91.7% (95%CI: 90.3%~93.9%)。Endosmart 对胃各部位整体识别的微平均 F1 分数为 94.6% (95%CI: 93.5%~95.6%),4 名中年资内镜医师为 92.1% (95%CI: 90.6%~93.5%),二者识别效能差异具有统计学意义 ($P < 0.001$)。结论 利用 ResNeSt-50 网络构建的 Endosmart 多标签分类模型在胃镜图像解剖部位识别中表现良好,识别效能优于中年资内镜医师。该模型可作为胃镜检查的辅助工具,用于检测胃镜检查部位的完整性,降低盲点率,提高胃镜检查质量。

【关键词】 胃镜检查; 人工智能; 深度学习; 多标签分类; 盲区监测

基金项目:国家自然科学基金(82370545);天津市科技计划项目(24ZYCGSY00900)

Construction and validation of a multi-label classification artificial intelligence model for assessing the completeness of gastroscopy

Lin Lin^{1,2}, Yao Shuangzhe¹, Song Yan¹, Xu Xin¹, Hui Yangyang¹, Wang Yufeng³, Feng Yue³, Mu Jinbao³, Lu Yi⁴, Chen Xin¹, Wang Bangmao¹, Zheng Zhongqing¹

¹Department of Gastroenterology, Tianjin Medical University General Hospital, Tianjin 300052, China;

²Department of Gastroenterology, Tianjin Medical University General Hospital Airport Hospital, Tianjin 300308, China;

³Tianjin Yujin Artificial Intelligence Medical Technology Co., Ltd, Tianjin 300384, China;

DOI: 10.3760/cma.j.cn321463-20250623-00127

收稿日期 2025-06-23 本文编辑 许文立

引用本文:林琳,姚双喆,宋岩,等.基于多标签分类的胃镜检查完整度评估人工智能模型的构建与验证[J].

中华消化内镜杂志,2026,43(5):365-373. DOI: 10.3760/cma.j.cn321463-20250623-00127.



⁴Department of Gastrointestinal Endoscopy, The Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou 510655, China

Corresponding author: Wang Bangmao, Email: mwang02@tmu.edu.cn; Zheng Zhongqing, Email: zzheng01@tmu.edu.cn.

【Abstract】 Objective To evaluate the application value of a multi-label classification model (Endosmart) based on ResNeSt-50 network for gastroscopic site recognition. **Methods** A total of 10 172 gastroscopic images involving 24 gastric regions and 13 617 labels were enrolled from Tianjin Medical University General Hospital Airport Hospital and Tianjin Nankai Hospital from January 2018 to July 2022. Of these, 8 501 images were used to develop the Endosmart via deep learning, enabling simultaneous multi-label prediction of multiple anatomical sites from a single gastroscopic image. Internal validation was performed using 1 671 gastroscopic images. External validation was conducted using 100 gastroscopic video clips from Tianjin Medical University General Hospital from August 2022 to December 2024. The precision, recall, and F1-score of Endosmart for gastric site recognition were assessed in both the internal image and external video validation sets. The overall recognition performance of Endosmart was further compared with those of 4 mid-career endoscopists in the video validation set. **Results** In the internal image validation set, the Endosmart achieved an average precision of 94.2% (95%CI: 92.6%-95.5%), a recall of 90.3% (95%CI: 88.7%-91.7%), and an F1-score of 86.2% (95%CI: 84.2%-88.2%) for gastric anatomical site recognition at the image level. In the external video validation set, the precision, recall and macro-average F1-score of Endosmart for comprehensive recognition of gastric anatomical sites were 95.5% (95%CI: 94.0%-96.9%), 93.7% (95%CI: 92.4%-95.0%) and 94.5% (95%CI: 93.3%-95.6%), respectively. While the 4 mid-career endoscopists yielded an average precision of 95.5% (95%CI: 93.9%-96.9%), a recall of 90.7% (95%CI: 89.1%-91.8%), and a macro-average F1-score of 91.7% (95%CI: 90.3%-93.9%). In the gastric anatomical site recognition task, the per-video micro F1-score of the Endosmart was 94.6% (95%CI: 93.5%-95.6%), compared with the average level of 92.1% (95%CI: 90.6%-93.5%) among the 4 endoscopists. The difference in recognition performance between the model and endoscopists was statistically significant ($P < 0.001$). **Conclusion** The ResNeSt-50-based Endosmart multi-label classification model demonstrates excellent performance in the recognition of anatomical sites from gastroscopic images and videos, and shows superior accuracy to those of mid-career endoscopists. This model can serve as an auxiliary tool for gastroscopy to monitor examination completeness, reduce blind spots, and improve examination quality.

【Key words】 Gastroscopy; Artificial intelligence; Deep learning; Multi-label classification; Blind spot monitoring

Fund program: National Natural Science Foundation of China (82370545); Tianjin Science and Technology Plan Project (24ZYCGSY00900)

胃镜检查是诊断上消化道疾病的重要工具,其检查质量直接影响疾病的发现和诊断^[1-3]。尽管放大内镜、色素内镜及窄带光成像等辅助诊断技术在临床中的应用日益广泛,但白光胃镜检查仍是高质量胃镜检查的基础和前提。然而,内镜医师操作经验的差异性、工作疲劳、检查的速度、视野盲区、设备新旧等均可能导致胃镜检查中病变遗漏^[2,4]。胃镜检查质量的参差对患者的健康保证、医疗资源的合理应用产生重大影响。

近年来,人工智能技术快速发展,在消化道内镜检查领域的应用显示出巨大潜力^[5]。开发基于人工智能的实时胃部解剖部位识别及检查完整性评估系统,可有效提升内镜医师胃镜检查操作过程中的规范性和完整性^[6-10]。使用人工智能胃部识别系统能够显著降低胃镜检查的盲点率,并延长检查时间,提高记录的完整性^[8,11]。然而,以往人工智能对胃内解剖部位的识别算法均采用单标签分类

框架^[7-16],对胃镜图片进行单部位分类,无法对存在多部位的胃镜图片进行同时识别及分析,计算效能和数据利用率较低。本研究旨在构建基于ResNeSt-50网络的多标签分类模型(下文称为Endosmart),以实现胃镜图像中胃解剖部位的精准识别与检查完整度评估。

资料与方法

一、图像纳入与排除标准

本研究收集了天津医科大学总医院空港医院及天津市南开医院2018年1月至2022年7月间的上消化道内镜检查图片数据。根据排除标准排除不达标的图片数据后,获得符合质量要求的胃部图片10 172张,按8:2比例随机分配为训练集(8 501张)和验证集(1 671张),其中训练集用于模型构建,验证集用于内部验证。排除标准包括:(1)胃

部准备差,如存在较多黏液、气泡、重影、散光、模糊、出血;(2)患者年龄<18岁;(3)胃大部切除术后或胃全切术后,失去完整胃各部分结构者;(4)胃镜设备老旧导致的画面模糊、光线过暗无法观察胃各部位。图片均在白光模式下拍摄,使用日本富士或日本奥林巴斯公司胃镜。完成数据清洗和治理后,由2名具有5年以上内镜操作经验的内镜医师分别对10 172张合格图像进行独立标注和复核。若出现意见分歧,则由具有10年以上内镜操作经验的专家进行仲裁,待达成一致意见后,数据方可入库。因此,本研究重点针对胃内解剖部位的识别展开分析,以提高结果的针对性。

在数据标注阶段,考虑到胃镜检查图片的特殊性,即单张图片可能同时包含多个解剖部位(如胃窦大弯与前壁的重叠区域),本研究采用多标签标注方式。结合日本胃镜检查部位22分类、欧洲胃镜检查部位10分类以及本研究中人工智能系统的算法特点,将胃镜图片分为贲门(小弯侧、前壁、后壁)、胃底(穹窿部、小弯侧)、胃体上部(前壁、后壁、小弯)、胃体中部(前壁、后壁、小弯、中上部大弯)、胃体下部(前壁、后壁、小弯、大弯)、胃角(前壁、后壁、小弯)、胃窦(前壁、后壁、小弯、大弯)、幽门共24个部位标签。本研究共纳入10 172张图片,包括13 617个标签。

收集天津医科大学总医院2022年8月至2024年12月间的100部胃镜检查视频资料作为外部验证。4名具有3年以上内镜操作经验的中年资

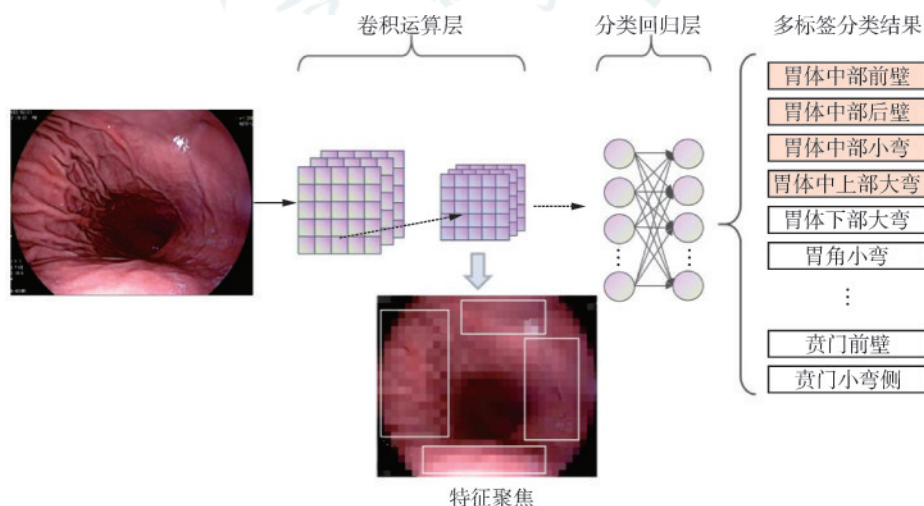
内镜医师进行独立阅片评估,统计胃镜检查视频中各解剖部位的观察完整性,并详细记录已观察及未观察到的具体部位。

二、模型的构建

在算法设计方面,本研究基于ResNeSt-50网络构建多标签分类模型Endosmart^[17-18](图1)。该模型采用注意力拆分(split-attention)结构,结合优化的残差连接设计和自适应特征聚合机制,有效提升了对胃镜图像中细微解剖特征的捕捉能力。在网络输出层,通过Sigmoid激活函数将特征向量转换为多个二分类任务,并采用二元交叉熵损失函数(BCELoss)进行模型训练。在推理阶段,系统对每个类别的输出概率设置0.5的判别阈值,实现胃部位置的多标签分类。图2展示了该模型的特征可视化结果。

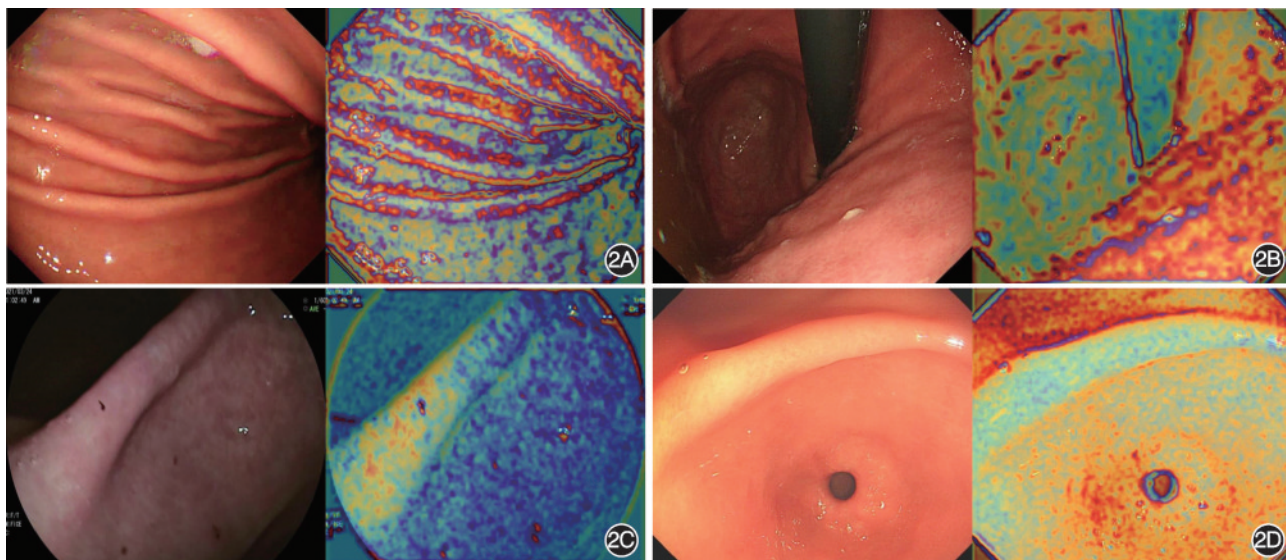
三、研究评价指标及统计学方法

本研究通过应用Python进行统计分析。以精确度(precision)、召回率(recall)和F1分数(F1 score)作为内部图片验证、外部视频验证的核心评价指标。其中,精确度又称阳性预测值(positive predictive value, PPV),反映在所有预测为阳性的结果中,真实为阳性的比例;召回率即灵敏度(sensitivity),反映在所有实际为阳性的样本中,被正确预测为阳性的比例,提示模型对某一特定部位的检出能力;而F1分数则作为精确度与召回率的调和平均值用于综合评估模型性能, $F1 \text{ 分数} = 2 \times (\text{精确度} \times \text{召回率}) / (\text{精确度} + \text{召回率})$ 。



注:本研究设计的Endosmart基于ResNeSt-50网络提取胃镜图像特征。ResNeSt-50的注意力拆分结构能有效捕捉不同胃部位置的细节特征,如胃体的褶皱、胃角的形态等。在网络的输出层,应用Sigmoid激活函数处理特征向量,将多标签分类问题转化为多个二分类任务,并使用二元交叉熵损失函数(BCELoss)进行模型训练。推理阶段,系统对每个类别的输出概率设定阈值(0.5),超过阈值的标签被判定为存在,从而实现胃镜图像的多标签分类

图1 Endosmart算法设计示意图



注:原始内镜图像(左)与 Endosmart 生成的注意力图(右),注意力图中红色区域表示模型重点关注的解剖结构特征,蓝色区域代表较低关注度

图 2 Endosmart 模型下胃镜图像识别过程的可视化 2A:胃体大弯,Endosmart 关注大弯皱襞的特征;2B:胃底前壁、小弯,Endosmart 关注镜头高亮聚焦区域,对应胃底前壁黏膜与旁侧小弯黏膜;2C:胃角小弯侧、胃角前壁,Endosmart 关注胃角解剖结构尤其是图左下侧前壁区;2D:胃窦小弯、幽门,Endosmart 关注了胃窦腔体结构及幽门

外部视频验证集中,以 2 名从业 10 年以上资深内镜专家的一致判读结果作为金标准。分别评估 Endosmart 及 4 名中年资内镜医师的独立识别效能,计算精确度、召回率、F1 分数及 95%CI;同时整合 4 名医师判读结果取均值,并同步测算各项评价指标及 95%CI。本研究设置两个统计分析维度完成模型与医师的效能对比。第一部分为整体效能对比,从部位层面和视频层面分别展开。(1)部位层面:以 24 个胃部解剖标签部位为分析单元,分别计算各部位的精确度、召回率及 F1 分数,再以宏平均方式汇总宏平均 F1 分数,通过对 100 例视频进行 Bootstrap 重采样(10 000 次)获得各指标的 95%CI。(2)视频层面:以单例胃镜视频为分析单元,将该视频内 24 个部位的真阳性、假阳性及假阴性数据加总后计算 Endosmart 及内镜医师平均水平的微平均 F1 分数,通过对 100 例视频进行 Bootstrap 重采样(10 000 次)获得 Endosmart 及内镜医师微平均 F1 分数的 95%CI,并采用配对置换检验(50 000 次)完成模型与内镜医师微平均 F1 分数的差异分析。第二部分为单部位效能对比,分别计算 Endosmart 与内镜医师平均水平在各个胃部解剖标签部位的 F1 分数,通过 Bootstrap 重采样(10 000 次)获得各部位 F1 分数的 95%CI,并采用配对置换检验(50 000 次)对各部位 F1 分数进行统计学差异分析。所有统计学检验均为双侧检验, $P < 0.05$ 为差异有统计学意义。

结 果

一、内部图片验证结果

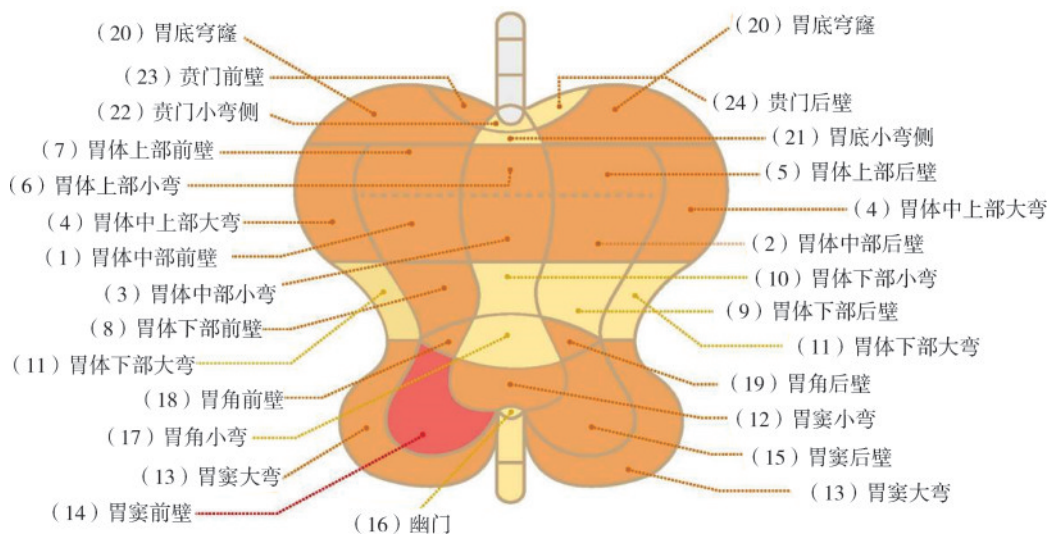
Endosmart 在胃镜图片中识别胃各部位的平均精确度为 94.2% (95%CI: 92.6%~95.5%),召回率为 90.3% (95%CI: 88.7%~91.7%),F1 分数为 86.2% (95%CI: 84.2%~88.2%),具体结果见表 1。Endosmart 在贲门小弯侧(F1=93.1%, 95%CI: 91.4%~94.6%)、贲门后壁(F1=93.4%, 95%CI: 91.8%~94.8%)、胃底小弯侧(F1=96.3%, 95%CI: 94.9%~97.4%)、胃体下部后壁(F1=93.2%, 95%CI: 91.6%~94.6%)、胃体下部小弯(F1=95.4%, 95%CI: 94.0%~96.6%)等易出现视野盲区的部位的 F1 分数均超过 90.0%;除了胃窦前壁(F1=75.3%, 95%CI: 72.5%~77.9%),其余部位的 F1 分数均保持在 80.0% 以上(图 3)。

二、外部视频验证结果

Endosmart 在胃镜视频验证集中对胃镜下胃各部位的整体识别精确度为 95.5% (95%CI: 94.0%~96.9%),召回率为 93.7% (95%CI: 92.4%~95.0%),宏平均 F1 分数为 94.5% (95%CI: 93.3%~95.6%);4 名中年资内镜医师对胃镜下胃各部位整体识别的平均精确度为 95.5% (95%CI: 93.9%~96.9%),召回率为 90.7% (95%CI: 89.1%~91.8%),宏平均 F1 分数为 91.7% (95%CI: 90.3%~93.9%)。Endosmart 对

表 1 Endosmart 在图片内部验证中对胃各部位的识别结果

部位	精确度(95%CI)	召回率(95%CI)	F1 分数(95%CI)
贲门小弯侧	91.2%(89.2%~93.0%)	95.8%(94.2%~97.0%)	93.1%(91.4%~94.6%)
贲门前壁	83.1%(80.3%~85.7%)	94.9%(93.2%~96.2%)	88.9%(86.8%~90.8%)
贲门后壁	96.9%(95.5%~98.0%)	89.2%(87.0%~91.2%)	93.4%(91.8%~94.8%)
胃底穹窿	74.8%(71.6%~77.8%)	86.3%(83.7%~88.5%)	80.2%(77.6%~82.6%)
胃底小弯侧	97.1%(95.7%~98.2%)	94.8%(93.2%~96.1%)	96.3%(94.9%~97.4%)
胃体上部前壁	94.2%(92.4%~95.7%)	71.8%(68.6%~74.8%)	81.3%(78.7%~83.7%)
胃体上部后壁	87.3%(84.9%~89.5%)	80.9%(78.3%~83.3%)	84.1%(81.9%~86.2%)
胃体上部小弯	77.9%(75.1%~80.5%)	82.6%(80.0%~85.0%)	80.4%(77.8%~82.8%)
胃体中部前壁	97.8%(96.6%~98.7%)	86.2%(83.8%~88.4%)	85.4%(83.0%~87.6%)
胃体中部后壁	91.3%(89.3%~93.1%)	93.7%(92.1%~95.1%)	86.2%(84.0%~88.4%)
胃体中部小弯	95.9%(94.5%~97.0%)	88.6%(86.4%~90.6%)	85.3%(82.9%~87.5%)
胃体中上部大弯	96.8%(95.4%~97.9%)	86.4%(84.0%~88.6%)	86.2%(84.0%~88.4%)
胃体下部前壁	96.1%(94.7%~97.2%)	78.3%(75.5%~80.9%)	86.4%(84.2%~88.6%)
胃体下部后壁	91.4%(89.4%~93.2%)	94.6%(93.0%~95.8%)	93.2%(91.6%~94.6%)
胃体下部小弯	93.6%(91.8%~95.1%)	96.8%(95.6%~97.8%)	95.4%(94.0%~96.6%)
胃体下部大弯	97.2%(96.0%~98.1%)	87.9%(85.7%~89.9%)	92.1%(90.3%~93.7%)
胃角小弯	92.4%(90.4%~94.1%)	88.3%(86.1%~90.3%)	90.2%(88.2%~92.0%)
胃角前壁	88.6%(86.2%~90.8%)	74.8%(71.8%~77.6%)	81.5%(78.9%~83.9%)
胃角后壁	85.3%(82.7%~87.7%)	93.4%(91.8%~94.8%)	88.7%(86.5%~90.7%)
胃窦前壁	86.4%(83.8%~88.6%)	67.2%(64.0%~70.2%)	75.3%(72.5%~77.9%)
胃窦后壁	85.6%(83.0%~87.8%)	84.3%(81.9%~86.5%)	85.1%(82.7%~87.3%)
胃窦小弯	86.2%(83.8%~88.4%)	79.4%(76.8%~81.8%)	82.3%(79.9%~84.5%)
胃窦大弯	88.7%(86.3%~90.9%)	77.3%(74.5%~79.9%)	82.5%(80.1%~84.7%)
幽门	85.3%(82.9%~87.5%)	97.8%(96.8%~98.6%)	90.9%(89.1%~92.5%)



注:黄色代表在内部验证集中模型的F1分数超过90.0%;橙色代表在内部验证集中模型的F1分数在80.0%~90.0%;红色代表在内部验证集中模型的F1分数低于80.0%

图 3 Endosmart 模型的 24 个标签及在图片内部验证中各部位的 F1 分数示意图

于胃镜下各部位整体识别的微平均 F1 分数优于中年资内镜医师平均水平 (94.6% 比 92.1%, $P < 0.001$), 详细结果见表 2。

在 24 个具体解剖部位识别中 (表 3、4), Endosmart 在 21 个部位的 F1 分数均超过 90.0%。在贲门(小弯侧、前壁、后壁)、胃体上部及中部后

表 2 Endosmart 与 4 名中年资内镜医师对视频外部验证集中胃各部位的整体识别结果

识别者	精确度(95%CI)	召回率(95%CI)	宏平均 F1 分数(95%CI)	微平均 F1 分数(95%CI)
Endosmart	95.5%(94.0%~96.9%)	93.7%(92.4%~95.0%)	94.5%(93.3%~95.6%)	94.6%(93.5%~95.6%)
医师平均水平	95.5%(93.9%~96.8%)	90.5%(89.1%~91.8%)	91.7%(90.3%~92.9%)	92.1%(90.6%~93.5%)
医师 1	97.7%(96.6%~98.8%)	82.4%(78.8%~85.7%)	89.0%(86.5%~91.2%)	—
医师 2	97.4%(96.0%~98.5%)	83.4%(81.7%~85.0%)	86.8%(85.3%~88.1%)	—
医师 3	94.6%(92.7%~96.3%)	97.0%(95.2%~98.4%)	95.7%(94.2%~97.0%)	—
医师 4	92.1%(89.8%~94.2%)	99.3%(98.8%~99.7%)	95.4%(93.9%~96.6%)	—

注：“—”指无相关数据

表 3 Endosmart 与中年资内镜医师在视频外部验证集中识别胃各解剖部位的精确度和召回率

部位	Endosmart		医师平均水平	
	精确度(95%CI)	召回率(95%CI)	精确度(95%CI)	召回率(95%CI)
贲门小弯侧	80.3%(70.0%~87.7%)	89.7%(80.2%~94.9%)	88.4%(83.6%~92.6%)	70.6%(67.3%~73.8%)
贲门前壁	89.0%(80.4%~94.1%)	91.2%(83.0%~95.7%)	92.4%(88.9%~95.7%)	70.9%(67.4%~74.2%)
贲门后壁	82.4%(72.9%~89.0%)	95.9%(88.6%~98.6%)	89.0%(84.4%~93.0%)	71.9%(68.9%~75.0%)
胃底穹窿	98.9%(93.8%~99.8%)	87.8%(79.8%~92.9%)	98.5%(96.2%~100.0%)	98.2%(96.6%~99.5%)
胃底小弯侧	75.8%(64.2%~84.5%)	64.9%(53.8%~74.7%)	88.2%(82.2%~93.4%)	92.5%(89.1%~95.7%)
胃体上部前壁	98.9%(94.3%~99.8%)	97.9%(92.7%~99.4%)	98.5%(96.6%~99.8%)	94.0%(90.7%~96.9%)
胃体上部后壁	100.0%(95.9%~100.0%)	97.8%(92.5%~99.4%)	95.5%(91.8%~98.5%)	91.4%(87.6%~94.7%)
胃体上部小弯	97.8%(92.4%~99.4%)	94.7%(88.3%~97.7%)	98.1%(95.9%~99.7%)	90.8%(87.1%~94.1%)
胃体中部前壁	98.9%(94.3%~99.8%)	97.9%(92.7%~99.4%)	98.5%(96.6%~99.8%)	94.0%(90.7%~96.9%)
胃体中部后壁	100.0%(95.9%~100.0%)	97.8%(92.5%~99.4%)	95.5%(91.8%~98.5%)	91.4%(87.6%~94.7%)
胃体中部小弯	97.8%(92.4%~99.4%)	94.7%(88.3%~97.7%)	98.1%(95.9%~99.7%)	90.8%(87.1%~94.1%)
胃体中上部大弯	97.8%(92.4%~99.4%)	96.8%(90.9%~98.9%)	95.7%(92.0%~98.7%)	93.5%(90.4%~96.3%)
胃体下部前壁	100.0%(96.0%~100.0%)	95.9%(89.9%~98.4%)	97.6%(94.4%~100.0%)	91.8%(88.3%~94.8%)
胃体下部后壁	97.7%(91.9%~99.4%)	87.5%(79.4%~92.7%)	96.8%(93.6%~99.2%)	87.5%(83.5%~91.2%)
胃体下部小弯	100.0%(95.9%~100.0%)	91.9%(84.9%~95.8%)	98.9%(96.5%~100.0%)	88.9%(85.1%~92.5%)
胃体下部大弯	100.0%(95.9%~100.0%)	91.8%(84.7%~95.8%)	98.2%(95.5%~100.0%)	100.0%(100.0%~100.0%)
胃角小弯	98.9%(94.3%~99.8%)	96.9%(91.3%~98.9%)	98.7%(96.3%~100.0%)	96.6%(94.3%~98.7%)
胃角前壁	92.9%(85.3%~96.7%)	98.7%(93.2%~99.8%)	87.3%(81.0%~92.8%)	85.4%(81.0%~89.7%)
胃角后壁	86.2%(77.0%~92.1%)	95.8%(88.5%~98.6%)	82.3%(74.8%~89.1%)	84.7%(79.7%~89.6%)
胃窦前壁	99.0%(94.3%~99.8%)	96.9%(91.4%~99.0%)	98.2%(95.3%~100.0%)	96.2%(93.6%~98.2%)
胃窦后壁	100.0%(96.0%~100.0%)	95.8%(89.8%~98.4%)	96.8%(93.6%~99.2%)	95.1%(92.4%~97.4%)
胃窦小弯	100.0%(96.0%~100.0%)	93.0%(86.3%~96.6%)	100.0%(100.0%~100.0%)	97.8%(95.8%~99.2%)
胃窦大弯	100.0%(96.3%~100.0%)	99.0%(94.6%~99.8%)	100.0%(100.0%~100.0%)	98.5%(96.8%~99.8%)
幽门	100.0%(96.3%~100.0%)	99.0%(94.6%~99.8%)	100.0%(100.0%~100.0%)	99.8%(99.2%~100.0%)

壁、胃角前壁与后壁, Endosmart 的 F1 分数显著高于中年资内镜医师平均水平($P < 0.05$)。值得注意的是, 中年资内镜医师在胃底(穹窿、小弯侧)、胃体下部大弯的 F1 分数优于 Endosmart($P < 0.05$)。其余各部位 Endosmart 与中年资内镜医师平均水平的 F1 分数差异无统计学意义($P > 0.05$)。

讨 论

本研究旨在评估多标签分类人工智能模型 Endosmart 在胃镜检查中对胃各部位识别及盲区监

测的应用价值。本研究结果显示, 在内部图片验证中, Endosmart 对胃各部位识别的平均精确度为 94.2% (95%CI: 92.6%~95.5%), 召回率为 90.3% (95%CI: 88.7%~91.7%), F1 分数为 86.2% (95%CI: 84.2%~88.2%); 在外部视频验证中, 该模型展现出稳健的泛化性能, 对胃各部位综合识别的精确度、召回率及宏平均 F1 分数分别为 95.5% (95%CI: 94.0%~96.9%)、93.7% (95%CI: 92.4%~95.0%) 和 94.5% (95%CI: 93.3%~95.6%)。Endosmart 在胃镜检查中对胃各解剖部位的整体识别效能优于中年资内镜医师平均水平, 可作为内镜检查的辅助识别

表 4 Endosmart 与中年资内镜医师在视频外部验证集中识别胃各解剖部位的 F1 分数(95%CI)及比较

部位	Endosmart	医师平均水平	P 值
贲门小弯侧	84.7%(77.8%~90.5%)	70.8%(66.2%~74.7%)	0.005
贲门前壁	90.1%(84.8%~94.5%)	73.6%(69.6%~77.1%)	<0.001
贲门后壁	88.6%(82.9%~93.4%)	72.6%(68.3%~76.4%)	<0.001
胃底穹窿	93.0%(88.8%~96.4%)	98.3%(96.9%~99.5%)	0.010
胃底小弯侧	69.9%(60.7%~77.9%)	89.9%(86.0%~93.2%)	<0.001
胃体上部前壁	98.4%(96.3%~100.0%)	96.0%(94.0%~97.8%)	0.127
胃体上部后壁	98.9%(97.2%~100.0%)	93.1%(90.2%~95.5%)	0.003
胃体上部小弯	96.3%(93.3%~98.9%)	94.1%(91.6%~96.2%)	0.295
胃体中部前壁	98.4%(96.3%~100.0%)	96.0%(94.0%~97.8%)	0.128
胃体中部后壁	98.9%(97.2%~100.0%)	93.1%(90.2%~95.5%)	0.003
胃体中部小弯	96.3%(93.3%~98.9%)	94.1%(91.6%~96.2%)	0.295
胃体中上部大弯	97.3%(94.6%~99.5%)	94.5%(92.0%~96.7%)	0.142
胃体下部前壁	97.9%(95.7%~99.5%)	94.3%(91.7%~96.4%)	0.078
胃体下部后壁	92.3%(88.0%~95.8%)	91.2%(88.1%~93.9%)	0.724
胃体下部小弯	95.8%(92.5%~98.5%)	93.3%(90.7%~95.6%)	0.269
胃体下部大弯	95.7%(92.4%~98.4%)	99.1%(97.7%~100.0%)	0.025
胃角小弯	97.9%(95.7%~99.5%)	97.6%(95.9%~99.0%)	0.856
胃角前壁	95.7%(92.2%~98.6%)	85.5%(81.1%~89.4%)	<0.001
胃角后壁	90.8%(85.5%~95.1%)	82.8%(77.7%~87.3%)	0.024
胃窦前壁	97.9%(95.7%~99.5%)	97.1%(95.1%~98.7%)	0.576
胃窦后壁	97.9%(95.6%~99.5%)	95.8%(93.6%~97.7%)	0.257
胃窦小弯	96.4%(93.6%~99.0%)	98.8%(97.8%~99.6%)	0.141
胃窦大弯	99.5%(98.5%~100.0%)	99.2%(98.3%~99.9%)	0.766
幽门	99.5%(98.5%~100.0%)	99.9%(99.6%~100.0%)	0.875

注:采用 50 000 次配对置换检验对 Endosmart 与内镜医师识别胃各解剖部位的 F1 分数进行组间比较

工具,为规范部位观察、减少漏检提供技术支撑。

胃镜检查过程中漏诊率较高,我国胃癌发现时多为进展期,早期胃癌检出率显著低于日本(80%)和韩国(50%)的检出水平^[19-20]。胃镜漏诊与胃镜视野盲区、内镜医师经验及工作疲劳、胃镜设备等因素相关。一项纳入 22 项研究的荟萃分析结果显示,胃癌在胃镜检查中的漏诊率高达 9.4%;既往胃镜检查结果为阴性但需长期随访患者的研究中发现胃镜检查漏诊率为 10.0%;同时性胃癌的漏诊率最高达 23.3%^[21]。胃食管交界处、胃体是漏诊胃癌的高发部位^[21-22]。贲门周围、胃切线部位(前壁、后壁)以及倒镜观察时的胃体小弯侧和胃底贲门小弯侧等区域易存在检查盲区^[23-24]。胃体大弯病灶可能因充气不足导致黏膜暴露不充分而被遗漏^[25]。经验不足的内镜医师往往有更高的漏诊率^[4,22]。胃镜检查质量还受内镜医师疲劳程度影响^[3,4]。为降低漏诊风险,本研究开发基于多标签分类的人工智能胃镜辅助检查系统 Endosmart,用于实时提示内镜医师在胃镜检查过程中关注潜在检查盲区。

既往人工智能在胃镜检查中胃部位识别的算法均采用单标签分类模型,此类模型在胃镜部位的识别过程中表现出极强的准确率,模型的结构简单高效,数据标注成本低^[8-16,26]。但是单标签分类模型只能选择概率最高的类别,导致信息丢失;胃解剖部位是人为划分的区域,各部位之间存在过渡,单标签分类模型对于解剖结构过渡区域容易出现误判。本研究采用的多标签分类框架更加贴近临床实际检查需求,通过 Sigmoid 激活函数和二元交叉熵损失,允许每个类别独立预测,如同时标记“胃窦前壁”和“胃窦小弯”^[18],解决了胃镜检查中因部位重叠和交叉导致的识别困难;利用注意力拆分结构可增强内镜图像的特征提取能力,增强对细微解剖部位特征的敏感性;优化的残差连接和自适应特征聚合可以融合多尺度信息,减少因图像模糊或遮挡导致的误分类,提升深层网络对复杂特征的表达能力;多标签模型可通过调整阈值平衡假阳性与假阴性,而单标签模型在重叠区域可能被迫牺牲部分类别^[16]。在实际应用中,Endosmart 可进行实时视

频流分析,为内镜医师提供观察完整度评估、截图完成率统计以及胃镜操作站点导航等功能,辅助临床操作与诊断。

胃各解剖部位识别任务的 F1 分数结果提示,Endosmart 的识别效能优于中年资内镜医师人工判读水平。虽然在静态图片验证中,由于胃窦前壁区域与胃窦小弯侧及胃窦大弯侧等毗邻结构解剖特征相似,模型易出现区域误判现象,导致该部位 F1 分数相对降低。然而,基于视频数据集的动态观察分析表明,Endosmart 对胃各部位的识别整体效能优于中年资内镜医师。与先前研究结果相似^[1],本研究在视频验证中发现 Endosmart (F1 分数 84.7%~90.1%) 与内镜医师 (F1 分数 70.8%~73.6%) 对贲门前壁、后壁、小弯侧的识别效能均相对较低,但 Endosmart 仍优于内镜医师平均水平 ($P < 0.05$); 值得注意的是,内镜医师对胃底(穹窿部、小弯侧)、胃体下部大弯的识别效能优于 Endosmart ($P < 0.05$)。贲门、胃底解剖结构具有复杂性与特殊性,且倒镜操作场景下,镜身遮挡、黏膜分泌物干扰、内镜旋转角度、检查观察速度、视野光照条件等诸多主观、客观因素会共同干扰视野观察,这是内镜医师与人工智能识别模型诊断结果不佳的主要原因;而人工智能在训练阶段对胃部图像的采集条件要求更为严格,可能进一步影响其在视频识别中的泛化能力。因此,人工智能的内镜辅助诊断的最佳效能应建立在视野充足、图像清晰、黏膜洁净的观察基础之上。此外,贲门解剖范围较小,临床实践中易被内镜医师忽略或误判为胃底,这也解释了 Endosmart 在贲门各部位识别的 F1 分数及灵敏度(召回率)均优于内镜医师的原因。

本研究结果表明,多标签分类模型能够有效识别胃内各解剖部位,为提高胃镜检查过程中病灶发现率奠定了重要基础。然而,作为一项回顾性研究,本研究尚缺乏更全面的外部验证数据。我们将基于该多标签分类模型,开展多中心外部验证,讨论本模型在不同品牌胃镜设备中的诊断效能,以及对内镜医师胃镜诊断完整率的影响,以进一步评估模型的准确性、稳定性及临床实用性。同时,我们将整合病灶诊断模块、电子染色内镜模块等功能,为临床医师提供更全面的智能辅助支持,从而优化胃镜检查流程,提升诊断效能。

本研究构建的 Endosmart 胃镜部位识别系统在精确度、召回率、F1 分数等评价指标方面达到预期水平。该系统可作为上消化道内镜检查的辅助工

具,协助医师实时监测检查部位,有助于提高胃镜检查的完整性和质量,降低盲区遗漏率,提升检查规范性。

利益冲突 所有作者声明不存在利益冲突

志谢 本研究在软件开发与算法设计方面得到了天津市御锦人工智能医疗科技有限公司的技术支持与算法协助;本研究的部分模型训练及内部验证数据由天津市中西医结合医院(天津市南开医院)消化科提供

作者贡献声明 林琳:研究实施、数据整理、数据统计分析与结果处理、论文撰写;姚双喆、宋岩、徐昕、惠洋洋:临床资料收集与研究实施;王玉峰、冯月、穆金宝:算法设计、研究方案优化及研究方案执行;卢祎:数据统计分析与结果处理;陈鑫、王邦茂、郑忠青:研究方案设计、课题指导及研究质量把控

参 考 文 献

- [1] Amin A, Gilmour H, Graham L, et al. Gastric adenocarcinoma missed at endoscopy[J]. J R Coll Surg Edinb, 2002, 47(5): 681-684.
- [2] Ferrari C, Tadros M. Enhancing the quality of upper gastrointestinal endoscopy: current indicators and future trends [J]. Gastroenterology Insights, 2024, 15(1):1-18. DOI: 10.3390/gastroent15010001.
- [3] Chiu P. Editorial on quality standards in upper gastrointestinal endoscopy: a position statement of the British Society of Gastroenterology (BSG) and Association of Upper Gastrointestinal Surgeons of Great Britain and Ireland (AUGIS) [J]. Transl Gastroenterol Hepatol, 2018, 3:13. DOI: 10.21037/tgh.2018.01.05.
- [4] 程树红, 张志坚, 詹磊磊, 等. 早期胃癌与高级别上皮内瘤变内镜漏诊原因分析[J]. 中国内镜杂志, 2015, 21(9): 919-922.
- [5] Schmitz R, Werner R, Repici A, et al. Artificial intelligence in GI endoscopy: stumbling blocks, gold standards and the role of endoscopy societies[J]. Gut, 2022, 71(3): 451-454. DOI: 10.1136/gutjnl-2020-323115.
- [6] Chen D, Wu L, Li Y, et al. Comparing blind spots of unsedated ultrafine, sedated, and unsedated conventional gastroscopy with and without artificial intelligence: a prospective, single-blind, 3-parallel-group, randomized, single-center trial[J]. Gastrointest Endosc, 2020, 91(2): 332-339.e3. DOI: 10.1016/j.gie.2019.09.016.
- [7] Choi SJ, Khan MA, Choi HS, et al. Development of artificial intelligence system for quality control of photo documentation in esophagogastroduodenoscopy[J]. Surg Endosc, 2022, 36(1): 57-65. DOI: 10.1007/s00464-020-08236-6.
- [8] Wu L, Zhang J, Zhou W, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy[J]. Gut, 2019, 68(12): 2161-2169. DOI: 10.1136/gutjnl-2018-317366.
- [9] 李夏, 吴练练, 于红刚. 人工智能胃镜在盲区监测和自主图像采集中的应用研究[J]. 中华消化内镜杂志, 2019, 36(4): 240-245. DOI: 10.3760/cma.j.issn.1007-5232.2019.04.004.
- [10] Ding A, Li Y, Chen Q, et al. Gastric location classification during esophagogastroduodenoscopy using deep neural

networks[C]. 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, 2021.

[11] Wu L, He X, Liu M, et al. Evaluation of the effects of an artificial intelligence system on endoscopy quality and preliminary testing of its performance in detecting early gastric cancer: a randomized controlled trial[J]. *Endoscopy*, 2021, 53(12):1199-1207. DOI: 10.1055/a-1350-5583.

[12] Li YD, Zhu SW, Yu JP, et al. Intelligent detection endoscopic assistant: an artificial intelligence-based system for monitoring blind spots during esophagogastroduodenoscopy in real-time [J]. *Dig Liver Dis*, 2021, 53(2):216-223. DOI: 10.1016/j.dld.2020.11.017.

[13] Ahn BY, Lee J, Seol J, et al. Evaluation of an artificial intelligence-based system for real-time high-quality photodocumentation during esophagogastroduodenoscopy[J]. *Sci Rep*, 2025, 15(1): 4693. DOI: 10.1038/s41598-024-83721-9.

[14] Chang YY, Li PC, Chang RF, et al. Deep learning-based endoscopic anatomy classification: an accelerated approach for data preparation and model validation[J]. *Surg Endosc*, 2022, 36(6):3811-3821. DOI: 10.1007/s00464-021-08698-2.

[15] Dong Z, Wu L, Mu G, et al. A deep learning-based system for real-time image reporting during esophagogastroduodenoscopy: a multicenter study[J]. *Endoscopy*, 2022, 54(8): 771-777. DOI: 10.1055/a-1731-9535.

[16] 王士旭, 柯岩, 楚江涛, 等. 基于数据增强和混合神经网络的人工智能技术在上消化道内镜检查部位识别中的应用[J]. *中华消化内镜杂志*, 2023, 40(3):189-195. DOI: 10.3760/cma.j.cn321463-20211019-00628.

[17] Zhang H, Wu CR, Zhang ZY, et al. ResNeSt: split-attention networks[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2022.

[18] Yeh MC, Li YN. Multilabel deep visual-semantic embedding [J]. *IEEE Trans Pattern Anal Mach Intell*, 2020, 42(6): 1530-1536. DOI: 10.1109/TPAMI.2019.2911065.

[19] Yang K, Hu JK. Gastric cancer treatment: similarity and difference between China and Korea[J]. *Transl Gastroenterol Hepatol*, 2017, 2:36. DOI: 10.21037/tgh.2017.04.02.

[20] Yashima K, Shabana M, Kurumi H, et al. Gastric cancer screening in Japan: a narrative review[J]. *J Clin Med*, 2022, 11(15):4337. DOI: 10.3390/jcm11154337.

[21] Pimenta-Melo AR, Monteiro-Soares M, Libânio D, et al. Missing rate for gastric cancer during upper gastrointestinal endoscopy: a systematic review and meta-analysis[J]. *Eur J Gastroenterol Hepatol*, 2016, 28(9):1041-1049. DOI: 10.1097/MEG.0000000000000657.

[22] Ren W, Yu J, Zhang ZM, et al. Missed diagnosis of early gastric cancer or high-grade intraepithelial neoplasia[J]. *World J Gastroenterol*, 2013, 19(13): 2092-2096. DOI: 10.3748/wjg.v19.i13.2092.

[23] Sung IK, Kim YC, Yun JW, et al. Characteristics of advanced gastric cancer undetected on gastroscopy[J]. *Korean J Gastroenterol*, 2011, 57(5):288-293. DOI: 10.4166/kjg.2011.57.5.288.

[24] Suzuki H, Nonaka S, Maetani I, et al. Clinical and endoscopic features of metachronous gastric cancer with possible lymph node metastasis after endoscopic submucosal dissection and Helicobacter pylori eradication[J]. *Gastric Cancer*, 2023, 26(5):743-754. DOI: 10.1007/s10120-023-01394-1.

[25] Yao K. The endoscopic diagnosis of early gastric cancer[J]. *Ann Gastroenterol*, 2013, 26(1):11-22.

[26] Wu S, Chen Y, Li Z, et al. Towards multi-label classification: next step of machine learning for microbiome research[J]. *Comput Struct Biotechnol J*, 2021, 19: 2742-2749. DOI: 10.1016/j.csbj.2021.04.054.

• 读者 • 作者 • 编者 •

《中华消化内镜杂志》2026 年可直接使用英文缩写的常用词汇

ERCP(内镜逆行胰胆管造影术)	POEM(经口内镜食管下括约肌切开术)	Hb(血红蛋白)
EST(经内镜乳头括约肌切开术)	NOTES(经自然腔道内镜手术)	PaO ₂ (动脉血氧分压)
EUS(超声内镜检查术)	MRCP(磁共振胰胆管成像)	PaCO ₂ (动脉血二氧化碳分压)
EUS-FNA(超声内镜引导细针穿刺抽吸术)	GERD(胃食管反流病)	ALT(丙氨酸转氨酶)
EMR(内镜黏膜切除术)	RE(反流性食管炎)	AST(天冬氨酸转氨酶)
ESD(内镜黏膜下剥离术)	IBD(炎症性肠病)	AKP(碱性磷酸酶)
ENBD(经内镜鼻胆管引流术)	UC(溃疡性结肠炎)	IL(白细胞介素)
ERBD(经内镜胆道内支架放置术)	NSAIDs(非甾体抗炎药)	TNF(肿瘤坏死因子)
APC(氩离子凝固术)	PPI(质子泵抑制剂)	VEGF(血管内皮生长因子)
EVL(内镜下静脉曲张套扎术)	HBV(乙型肝炎病毒)	ELISA(酶联免疫吸附测定)
EIS(内镜下硬化剂注射术)	HBsAg(乙型肝炎病毒表面抗原)	RT-PCR(逆转录-聚合酶链反应)

(本刊编辑部)