基于荧光 ROSE 的幽门螺旋杆菌实时可视化诊断新方法：一项基于无金标准潜在类别分析的大样本对照研究

李炳辉 张林 侯艳红 吴凯 张静 杨汨

中国人民解放军总医院第八医学中心消化内科，北京 100091

通讯作者：张林，Email：stepinghuns2@163.com

摘要

目的 本研究旨在假定无金标准的条件下，比较荧光快速现场评估（Fluorescence Rapid On-Site Evaluation, F-ROSE）快速检测法、传统病理学观察法（苏木精 - 伊红染色 / 吉姆萨染色）、尿素呼气试验（Urea Breath Test, UBT）、快速尿素酶试验（Rapid urease test，RUT）及胃镜白光直视判断法对幽门螺旋杆菌（Helicobacter pylori, Hp）感染的诊断效能，结合大样本临床实测数据论证 F-ROSE 的诊断优势，以验证 F-ROSE 的临床应用价值。

方法 前瞻性纳入 317 例疑似 Hp 感染患者，在同次就诊周期内采用盲法同步进行 5 种方法检测：荧光快速现场评估（Fluorescence Rapid On-Site Evaluation, F-ROSE）快速检测法、传统病理学观察法（苏木精 - 伊红染色 / 吉姆萨染色）、尿素呼气试验（Urea Breath Test, UBT）、快速尿素酶试验（Rapid urease test，RUT）及胃镜白光直视判断法。采用无金标准评价框架进行分析：首先通过 Cohen’s Kappa 系数评估方法间的一致性；其次运用潜在类别分析（LCA）估算各方法的敏感性（Se）、特异性（Sp）及约登指数（J），结合受试者工作特征（ROC）曲线评价诊断效能，并引入贝叶斯敏感性分析验证 LCA 模型的稳健性。本研究框架无需预设临床金标准，依托多种非完美检测方法的联合检测反应模式挖掘潜在真实患病状态信息，通过统计模型反向推演真实分类情况，实现对各检测方法诊断效能客观、公平的对比评价。

结果经 LCA 模型估算，该人群潜在患病率约为 34.8%。按综合效能（约登指数）排序，五种方法依次为：F-ROSE（0.8672）> RUT（0.8069）> ^{13}C-UBT（0.7889）> 病理（0.6076）> 肉眼观察（0.6013）。其中，F-ROSE 敏感性最高（0.9472），漏诊率极低；病理检查特异性最高（0.9843），但敏感性偏低（0.6234）。贝叶

斯分析证实了 LCA 结果的高度稳健性。此外，F-ROSE 检测耗时仅 25–30 分钟，显著优于传统病理（24~72 小时）。

结论 综合本研究 317 例患者临床检测数据，F-ROSE 在五种 Hp 检测方法中整体诊断效能最佳，胃镜白光直视观察效能相对欠佳。F-ROSE 兼具高敏感性、短检测时长及优异的弱阳性样本识别能力，曲线下面积（AUC）在潜在类别软标签（riLCA ）框架下显著高于病理检查、^{13}C 呼气试验、及肉眼观察，与快速尿素酶试验表现接近、差异处于临界显著性水平；同时保持较高特异性，与常规方法一致性良好，临床落地性与实用性突出，有望成为 Hp 感染的临床快速检测首选手段。

关键词：荧光 ROSE 快速检测；幽门螺旋杆菌；传统病理学观察法；尿素呼气试验法；快速尿素酶试验；胃镜白光直视判断；大样本临床验证

A Novel Real-Time Visual Diagnostic Method for Helicobacter pylori Based on Fluorescence ROSE: A Large-Sample Controlled Study Using Latent Class Analysis Without a Gold Standard

Li Binghui, Zhang Lin, Hou Yanhong, Wu Kai, Zhang Jing, Yang Mi

Department of Gastroenterology, The Eighth Medical Center of Chinese PLA General Hospital, Beijing 100091, China

Corresponding author: Zhang Lin, Email: stepinghuns2@163.com

Abstract

Objective This study aimed to compare the diagnostic performance of five detection methods for Helicobacter pylori (Hp) infection—fluorescence rapid on-site evaluation (F-ROSE), conventional histopathological examination (hematoxylin-eosin staining/Giemsa staining), urea breath test (UBT), rapid urease test (RUT), and endoscopic white-light visual assessment—under the assumption of no gold standard.

10.12201/bmr.202606.00055V1

Using large-sample clinical data, we sought to demonstrate the diagnostic advantages of F-ROSE and validate its clinical utility.

Methods A total of 317 patients with suspected Hp infection were prospectively enrolled. All five tests were performed synchronously in a blinded manner during the same visit: F-ROSE, conventional histopathology (HE/Giemsa), UBT, RUT, and endoscopic white-light visual assessment. A no-gold-standard evaluation framework was adopted: (1) inter-method agreement was assessed using Cohen's kappa coefficient; (2) latent class analysis (LCA) was employed to estimate the sensitivity (Se), specificity (Sp), and Youden index (J) of each method, with diagnostic performance further evaluated by receiver operating characteristic (ROC) curves; and (3) Bayesian sensitivity analysis was introduced to verify the robustness of the LCA model. This framework did not require a prespecified clinical gold standard but instead leveraged the combined response patterns of multiple imperfect tests to infer the latent true infection status, enabling an objective and fair comparative evaluation of diagnostic performance across methods.

Results The LCA model estimated a latent prevalence of approximately 34.8% in this cohort. Ranked by overall efficacy (Youden index), the five methods were: F-ROSE (0.8672) > RUT (0.8069) > ^{13}C-UBT (0.7889) > histopathology (0.6076) > visual inspection (0.6013). F-ROSE exhibited the highest sensitivity (0.9472), indicating a very low missed-diagnosis rate, while histopathology showed the highest specificity (0.9843) but relatively low sensitivity (0.6234). Bayesian analysis confirmed the high robustness of the LCA results. In addition, the turnaround time for F-ROSE was only 25–30 minutes, significantly shorter than that for conventional histopathology (24–72 hours).

Conclusion Based on the clinical data from 317 patients in this study, F-ROSE demonstrated the best overall diagnostic performance among the five Hp detection

10.12201/bmr.202606.00055V1

methods, while endoscopic white-light visual assessment exhibited relatively poor efficacy. F-ROSE combines high sensitivity, short turnaround time, and excellent ability to identify weakly positive samples. Its area under the curve (AUC) under the soft-label latent class (riLCA) framework was significantly superior to that of histopathology, ^{13}C-UBT, and visual inspection, and was close to that of RUT with a marginally significant difference. It also maintained reasonably high specificity and showed good agreement with conventional methods. With strong clinical feasibility and practicality, F-ROSE holds promise as the preferred rapid testing method for clinical diagnosis of Hp infection.

Keywords: fluorescence ROSE rapid test; Helicobacter pylori; conventional histopathological examination; urea breath test; rapid urease test; endoscopic white-light visual assessment; large-sample clinical validation

一、引言

幽门螺旋杆菌（Helicobacter pylori，Hp）是定植于人体胃黏膜的核心致病菌，也是诱发慢性胃炎、消化性溃疡、胃黏膜萎缩、肠化生等多种胃部疾病的关键危险因素**错误！不能识别的开关参数。**，更是胃癌发生发展的一类重要可控诱因，已被世界卫生组织列为一类致癌原**错误!不能识别的开关参数。**。及时准确地检测出 Hp 感染对于疾病的诊断、治疗及预防具有关键意义。目前临床常用 Hp 检测方法主要包括组织病理学、尿素呼气试验（UBT）、快速尿素酶试验（RUT）及胃镜肉眼判断，虽各有优势，但均存在局限性：传统病理可靠但流程繁琐、耗时较长、对萎缩 / 肠化生样本灵敏度下降，结果依赖操作者经验；UBT 无创便捷，但易受胃黏膜状态、铋剂、PPI 及抗生素影响；RUT 出结果快，但易受药物、操作污染干扰；胃镜肉眼判断主观性强、一致性差**错误！不能识别的开关参数。错误！不能识别的开关参数。**。荧光快速现场评估（Fluorescence Rapid On-Site Evaluation，F-ROSE）快速检测技术是一种新兴的检测手段，其原理是通过荧光染色标记试剂与消化道脱落细胞中的细胞核、Hp 菌体及中性粒细胞特异性结合，借助荧光显微镜在特定激发波长下成像，可对 Hp、中性粒细胞及异常细胞进行快速识别和人工智能辅助诊断（AI-assisted diagnosis）**错误！不能识别的开关参数。**。该检

测方法已在肺泡灌洗液病原诊断、宫颈液基细胞学检查等领域获得临床认可**错误！不能识别的开关参数。错误！不能识别的开关参数。错误！不能识别的开关参数。**，针对 Hp 诊断亦已有相关探索研究**错误！不能识别的开关参数。错误！不能识别的开关参数。**。尽管 F-ROSE 在 Hp 检测中已有初步探索，但现有研究多依赖传统金标准进行效能验证，未充分考虑各方法间的系统性偏倚；且缺乏基于大样本的联合检测策略优化证据。本研究旨在填补上述空白，在无金标准框架下对 F-ROSE 进行客观评价。本研究依托 317 例大样本临床数据，将 F-ROSE 与传统病理学、UBT、RUT 及胃镜白光直视判断法进行同步对照，从阳性率、方法间一致性及临床适配性等维度评价其应用价值，为 F-ROSE 的临床推广提供循证医学证据。

二、材料与方法

2.1 研究对象

选取 2025 年 10 月至 2026 年 4 月期间在我院消化内科就诊、因上消化道症状拟行胃镜检查的疑似 Hp 感染患者 317 例。

纳入标准：① 年龄 18-75 岁；② 自愿接受胃镜检查并取胃黏膜活检；③ 在本中心行胃镜检查并同意参与本研究，同时完成 F-ROSE、病理学检查、UBT、RUT 四项检测。

排除标准：① 入组前 4 周内服用过抗生素、铋剂或质子泵抑制剂，或 2 周内服用过 H2 受体拮抗剂；② 有胃部手术史；③ 内镜下怀疑或确诊为胃癌，或病理提示广泛肠上皮化生 / 萎缩；④ 有严重凝血功能障碍或活动性上消化道出血；⑤ 合并严重心、肝、肾等系统性疾病或免疫缺陷病；⑥ 妊娠或哺乳期妇女；⑦ 无法配合完成任一检测项目或拒绝参与研究。

2.2 实验材料

胃黏膜活检标本采集器械（南京南微医学科技股份有限公司）、幽门螺旋杆菌荧光 ROSE 快速检测试剂盒（江苏诺禹生物科技有限公司）、传统病理学染色试剂（苏木精 - 伊红染色液、Giemsa 染色液等）、尿素呼气试验检测仪器及试剂（深圳市中核海得威生物科技有限公司）、AI 智能高清荧光显微镜分析系统（江苏诺禹生物科技有限公司）、快速尿素酶检测试剂盒（上海惠泰医疗科技有限公司）、光学显微镜、全自动荧光免疫扫描仪（江苏诺禹生物科技有限公司）；所有试剂均在有效期内使用，仪器均完成定期校准。

2.3 实验方法

2.3.1 Hp 荧光 ROSE 快速检测

本研究采用三步荧光染色法，依次使用活菌染色试剂、死菌染色试剂与 Hp 荧光染料对胃黏膜标本进行染色。其中：死菌染色试剂含死菌探针 EthD-III；活菌染色试剂含活菌探针 DMAO；Hp 荧光染料含荧光染料吖啶橙与细胞骨架蛋白β-actin 抗体**错误！不能识别的开关参数。**。

具体步骤如下：将标本分为实验孔与对照孔；向对照孔滴加死菌染色试剂，避光染色 5 min；继续向同一对照孔滴加活菌染色试剂，避光染色 2 min；向实验孔滴加 Hp 荧光染料，染色 30 s；冲洗后加盖玻片，置于全自动荧光免疫扫描仪（江苏诺高生物科技有限公司）载物台，选择 Hp 荧光波段（激发波长 460～490 nm，发射波长 510 nm），于 ×400 倍下观察。

判读标准：实验孔用于判断 Hp 感染状态：出现典型橙红色荧光、形态符合 S 形、弧形或螺旋状，判定为 Hp 阳性；无特异性荧光为阴性。对照孔用于判断 Hp 活菌 / 死菌状态：活菌呈现绿色荧光，死菌呈现橙红色荧光。该双孔染色体系既可用于 Hp 快速诊断，也可进一步为 Hp 活性评估及药敏检测提供依据。

2.3.2 传统病理学观察法

HE 染色为本中心病理科观察幽门螺旋杆菌的染色方法。于高倍镜下（ ×1000）观察 Hp 形态特征，若观察到呈弧形短杆状、逗点状或“S”形、“C”形、“海鸥状”的细菌，且颜色与周围组织明显不同（HE 染色下呈紫红色），则判定为 Hp 阳性。

2.3.3 尿素呼气试验法

患者需空腹或禁食至少 2 小时。口服^{13}C-尿素试剂（10-15 ml）后静坐 15-30 分钟，期间避免饮食与剧烈活动。静坐结束后，患者向集气瓶内平缓呼气直至指示剂变色。应用红外线吸收光谱仪检测收集气体中的^{13}CO2 含量，若检测值≥4 DOB（或按说明书临界值）则判定为 Hp 阳性。

2.3.4 快速尿素酶试验

于胃镜下钳取幽门前区（距幽门 5cm 以内）大弯侧胃黏膜组织，立即将其植入 RUT 试剂凝胶中。将标本插入凝胶 1-2 分钟后，试剂变成红色而且红晕扩散即为

幽门螺旋杆菌检测结果为阳性，如果不变色或仅黏膜周围略变橙色，则至 30 分钟继续观察，期间颜色变红且红晕扩散也应判为幽门螺旋杆菌检测结果为阳性，不变色为阴性；若仅黏膜周围略变橙色，红晕不扩散，则判为阴性。

2.3.5 胃镜白光直视判断法

由 2 名具有 5 年以上胃镜操作经验的医师独立操作，在胃镜下观察胃黏膜形态。若观察到患者胃黏膜出现充血、水肿、糜烂、红斑、痘疹样、地图样改变或黏液白浊等 Hp 感染典型表现，判定为 Hp 阳性；黏膜形态正常，无典型感染表现，判定为阴性；意见不一致时，由第三名高年资医师复核确定结果。

2.3.6 检测随机化与盲法设计

为降低检测顺序和操作者主观偏倚，采用随机数字表法分配检测顺序，确保各检测序列均匀分布；所有检测由不同人员独立完成，检测人员均不知晓其他方法的检测结果，严格实施盲法，减少系统误差。

2.3.7 检测结果补充分析

鉴于目前检测 Hp 方法均存在假阳性与假阴性，故本研究不设立金标准，在原有的五种检测方法（荧光 ROSE 快速检测、传统病理学观察法、尿素呼气试验法、快速尿素酶试验及胃镜白光直视判断法）基础结果上，采用潜在类别分析（LCA）和贝叶斯方法进行补充分析。在此基础上进行五种方法的 ROC 曲线绘制分析。

2.4 统计学分析

2.4.1 描述性统计与率的比较：采用频数和百分率（%）描述患者的人口学特征（性别、年龄）及五种检测方法的阳性率。采用配对 McNemar 检验（连续校正）比较两两方法间阳性检出率的差异。

2.4.2 一致性分析：为评估任意两种检测方法间的一致性，对 5 种方法进行两两组合（共 C(5,2) = 10 对），分别构建 2×2 列联表，计算 Cohen's Kappa 系数（κ）及其 95%置信区间；一致性强度判定参照 Landis 与 Koch（1977）标准**错误！不能识别的开关参数。：** K< 0.20 为一致性差，0.20 ≤ K< 0.40 为一般，0.40 ≤ K< 0.60 为中等，0.60 ≤ K< 0.80 为较强，K ≥ 0.80 为几乎完全一

致。

2.4.3 潜在类别分析（Latent Class Analysis, LCA）：

LCA 是本研究无金标准诊断评价的核心统计方法**错误！不能识别的开关参数。**。鉴于现有幽门螺旋杆菌检测方法均存在不可忽视的假阳性与假阴性，本研究不预设金标准，将幽门螺旋杆菌真实感染状态视为潜在类别，基于 5 种检测方法的联合反应模式反推真实状态并估计诊断效能，实现公平评价。

2.4.3.1 模型原理与核心假设

基本原理：将研究对象划分为互斥穷尽的潜在类别，检测结果受潜在感染状态驱动，同步估计人群患病率与各方法敏感性、特异性**错误！不能识别的开关参数。**。

核心假设：模型基于局部独立性假设构建。该假设为潜在类别模型的经典设定，但已有研究提示，若检测间存在条件相关，可能导致参数估计偏倚**错误！不能识别的开关参数。**。本研究将通过拟合优度检验与实际检测关联模式联合验证假设满足程度。模型设定：采用 2 类别模型（感染类、未感染类），待估 11 个核心参数，包括 1 个潜在患病率、5 种方法的敏感性与特异性，此类无金标准条件下的潜在类别参数化评价框架在临床诊断性试验中已有成熟的应用与论证**错误！不能识别的开关参数。**，类别数合理性后续验证。

2.4.3.2 数据预处理与模型拟合

数据预处理：5 种检测结果标准化为二分类（阴性 = 0，阳性 = 1），排除缺失数据后纳入 317 例。参数估计：采用 EM 算法，通过 R 4.5.2 的 poLCA 包实现，迭代 E 步与 M 步至收敛（$\varepsilon = 10^{-6}$），设置 60 次随机起始点避免局部最优。

E 步（期望步）：于当前轮次的参数估计值$\hat{\theta}$，根据贝叶斯定理计算患者属于感染类的后验概率，作为软标签，公式：$r_i = P(D^+ \mid x_i, \hat{\theta})$。其中$x_i$为患者 i 的 5 项检测结果向量。该后验概率 ri 即为患者真实感染状态的概率性软标签（soft label），保留了全部不确定性信息；M 步：以所有患者的后验概率 ri 为权重，重新估计模型的全部 11 个参数。其中，第 j 种方法的敏感性与特异性按下式计算：公式：$\widehat{Se_j} = \frac{\sum_{i=1}^{n} r_i t_{ij}}{\sum_{i=1}^{n} r_i}, \widehat{Sp_j} = \frac{\sum_{i=1}^{n} (1-r_i)(1-t_{ij})}{\sum_{i=1}^{n} (1-r_i)}$。其中$t_{ij} \in \{0,1\}$为患者 i 在第

j 种方法上的二值检测结果。潜在患病率π 与各方法的条件响应概率同步更新。

2.4.3.3 最优类别数选择

拟合 K=1~4 类别模型。模型选择不依赖单一统计指标，而是综合考量以下四个维度**错误！不能识别的开关参数。错误！不能识别的开关参数。错误！不能识别的开关参数。**：第一，信息准则（AIC、BIC、aBIC）；第二，绝对拟合优度（局部独立性检验 P 值）；第三，潜在类别的临床可解释性与参数稳定性；第四，研究估计目标的结构匹配性——本研究旨在估计各检测方法相对于真实感染状态的敏感性与特异性，Se/Sp 估计量在定义上要求潜变量为二分类（感染 vs 未感染）。最终类别数的确定基于上述四个维度的综合判断，详见结果 3.7.1 节。

2.4.3.4 不确定性量化与软标签的后续应用

不确定性量化：采用 Bootstrap 法（B=500）估计诊断指标 95% CI，约登指数 $J = Se + Sp - 1$ 评价综合效能。软标签应用：保留患者后验感染概率作为软标签，不做硬分类，用于后续 ROC 分析与联合检测评价。

2.4.4 贝叶斯敏感性分析：本节采用贝叶斯潜在类别模型（Bayesian Latent Class Model, BLCM） 开展敏感性分析，将文献证据以信息性先验纳入建模，一方面验证 2.4.3 中 MLE 估计结果的稳健性，另一方面对 F-ROSE 诊断性能进行更全面的后验推断**错误！不能识别的开关参数。**。

2.4.4.1 先验分布设定：11 个待估参数分两类设定先验：信息性先验（9 个参数）：基于 Meta 分析结果，用矩匹配法设定 Beta 分布，等效样本量上限设为 50，避免压制实测数据**错误！不能识别的开关参数。错误！不能识别的开关参数。错误！不能识别的开关参数。错误！不能识别的开关参数。**；非信息性先验（2 个参数）：F-ROSE 的敏感性与特异性设为 Beta (1,1)，完全由本研究数据主导（F-ROSE 作为新兴荧光快速诊断技术，目前已有探索性研究但样本量小、未做 meta 分析、不足以构建信息性先验。）。

2.4.4.2 模型实现与收敛诊断：通过 Gibbs 抽样实现，3 条马尔可夫链各迭代 20000 次，预热 5000 次，间隔抽样后保留 9000 个后验样本。以 Gelman-Rubin 统计量 <1.05、有效样本量> 1000 判定收敛**错误！不能识别的开关参数。**，结果以中位数与 95% HPDI 报告。

2.4.4.3 先验稳健性检验

设置 3 种先验方案对比：主分析方案、全非信息方案、强先验方案。若各参数后验估计差异 < 5%，判定结论稳健，采用 Python 3.12 实现。

2.4.5 ROC 曲线绘制：本研究针对 5 种检测方法构建三套互补 ROC 分析体系，分别采用 LCA 软标签、贝叶斯软标签作为真实感染状态参照，同时新增贝叶斯后验概率为评分、各方法二分类结果为参照的一致性 ROC，全面评估诊断区分能力与方法间一致性，全面评估诊断效能。加权 ROC：以软标签为参照，计算加权敏感性、特异性与 AUC，二分类方法用单一工作点，分级方法用多工作点计算；一致性 ROC：以贝叶斯后验概率为评分、各方法二分类结果为参照，评估判定一致性；置信区间与检验：Bootstrap 法（B=1000）估计 AUC 的 95% CI，结合后验概率评估方法差异，用 Python 与 R pROC 包绘制。

2.4.6 DeLong 检验：采用加权 DeLong 法两两比较 ROC 曲线，适配软标签加权框架，计算 AUC 方差与协方差后正态近似检验，双侧 α =0.05，不做多重比较校正**错误！不能识别的开关参数。**。

2.4.7 两两联合检测策略评价：在不同临床场景中为了打破单一诊断瓶颈，本研究评估了五种检测方法两两联合后的诊断效能。

对 5 种方法的 10 个组合，分别采用并联（OR）、串联（AND） 规则，构建 20 种联合方案，用 LCA 软标签（r_i^{LCA}）与贝叶斯软标签（r_i^{Bayes}）（0.5 为阈值二值化）平行评价（仅在联合检测策略的串联/并联工作点列联表计算时，为构造 2×2 表必须将软标签二值化，故采用 0.5 作为决策阈值；ROC 与 AUC 计算仍使用原始软标签加权法；$r_i \geq 0.5$ 判为 Hp 阳性，否则为阴性）。

2.4.7.1 ROC 曲线与曲线下面积：以 LCA 软标签（r_i^{LCA}）与贝叶斯软标签（r_i^{Bayes}）为参照，采用加权 ROC 分析。对每对联合检测，将两种方法的二分类结果纳入 Logistic 回归，以回归模型输出的连续预测概率作为评分变量绘制 ROC 曲线，采用 DeLong 法估计 AUC 及其 95% CI。

2.4.7.2 串联（AND）与并联（OR）工作点：计算串联、并联工作点的 Se、Sp、PPV、NPV、PLR、NLR、约登指数，确定最佳工作点；

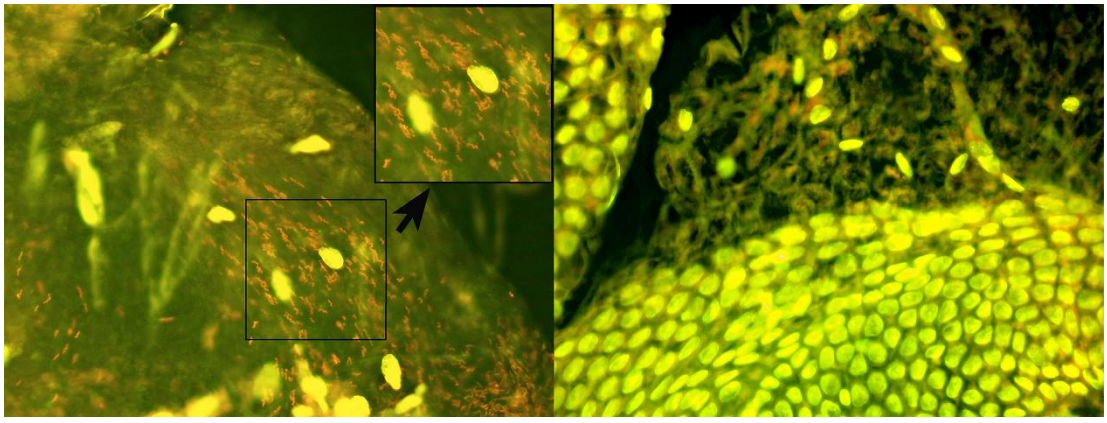
2.4.7.3 方案比较：配对 DeLong 检验比较联合方案 AUC 差异，P<0.05 为有统计学意义。

三、实验结果

3.1 荧光 ROSE 法与传统病理学方法 Hp 感染阳性结果染色特征

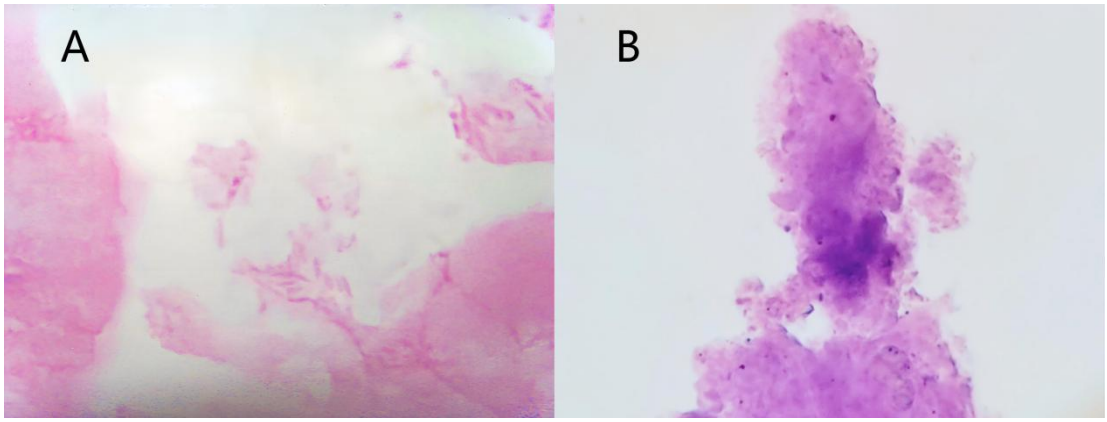
荧光 ROSE 采用实验孔与对照孔双孔染色体系。实验孔经 Hp 荧光染料染色后，Hp 阳性标本在暗背景下可见黄绿色荧光的细胞核，黏膜及黏液层中可见橙红色荧光（少数含活菌者活菌呈绿色荧光）的短棒状、弧状及卷曲状菌体（图 1 A），阴性标本无橙红色及绿色荧光菌体（图 1 B）；对照孔经活菌与死菌探针染色后，可区分活菌与死菌：活菌呈绿色荧光，死菌呈橙红色荧光（于 3.3 节详述）。传统病理学染色标本中，Hp 阳性标本于油镜（×1000）下可见黏膜上皮及黏液层中相应形态的菌体（图 2 A），阴性标本无染色菌体（图 2 B）。

图 1



注：Hp 荧光 ROSE 法染色结果（×400 倍）A：Hp 阳性标本，黏膜及黏液层可见橙红色荧光的短棒状、弧状及卷曲状菌体；B：Hp 阴性标本，无橙红色荧光菌体。

图 2



注：Hp 传统病理学 HE 染色结果（×1000 倍）A：Hp 阳性标本，黏膜及黏液层可见紫红色染色的短棒状、弧状及卷曲状菌体；B：Hp 阴性标本，无染色菌体。

3.2F-ROSE 阳性结果分级特征

F-ROSE 可依荧光信号强度与菌体数量进行阳性分级。阳性率以全部 40× 高倍复扫图片诊断出的活菌与死菌总数（Total）与阈值（Threshold）的关系计算，阈值当前设为 5 个（可按实际调整）。

具体的换算如下：

当 Threshold ＜ Total ≤ 2×Threshold 时， 结果记为 + （图 3A）；

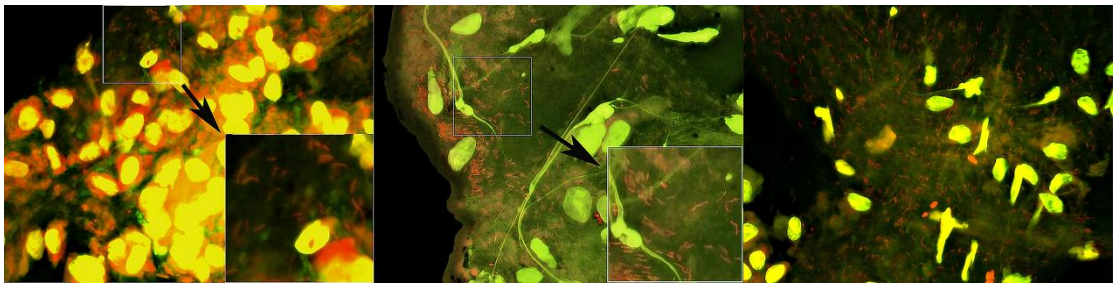
当 2×Threshold ＜ Total ≤ 4×Threshold 时， 结果记为 ++ (图 3B)；

当 Total ＞ 4×Threshold 时， 结果记为 +++ （图 3C）；

当 Total ≤ Threshold 时， Value = -；

本研究 317 例中，F-ROSE 阴性 196 例、阳性 121 例，其中“+”79 例、“++”22 例、“+++”20 例（图 4）。该分级分布提示 F-ROSE 可有效检出低载量感染病例；本研究未做分级与临床结局的关联，分级仅作定性参考，留待后续研究。除病理观察外，其他三种方法仅能定性判读，缺乏分级信息。

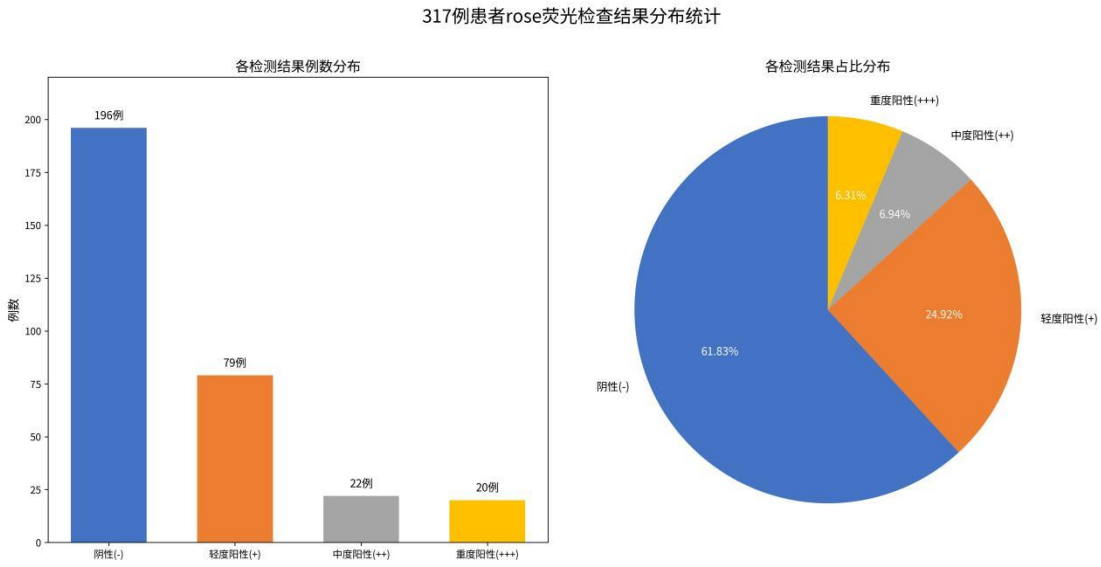
图 3



注：Hp 荧光 ROSE 法染色结果（×400 倍）A：检查结果为“+”；B：检查结果为“++”；C：检测结果为“+++”。

图 4

10.12201/bmr.202606.00055V1



注：A:F-ROSE 检测 4 类结果占比柱状图;B:F-ROSE 检测 4 类结果占比饼图。

3.3F-ROSE 阳性结果中鉴别死菌与活菌的特征

荧光染色过程中（图 5），活菌探针 DMAO 为膜通透性染料，可穿透完整活菌细胞膜并与核酸结合发绿色荧光；死菌探针 EthD-III 为膜非通透性染料，仅能穿过受损死菌的细胞膜并与核酸结合发红色荧光。成像结果中，活菌（图 6 绿色短杆状物）荧光较强、形态完整、轮廓清晰，可见典型杆状、弧形或 S 形结构；死菌（图 6 橙红色短杆状物）荧光中强、呈红/橘色，部分可见球化及碎片化。基于该原理可进一步实现 Hp 快速药敏检测**错误！不能识别的开关参数。**，本研究仅聚焦快速诊断，药敏检测将在后续研究中展开。121 例 F-ROSE 阳性样本中，观察到活菌者 23 例，均见于“++”或“+++”样本，且胃镜下均可见明显病变（多为红斑渗出、糜烂、结节样及萎缩性改变），提示活菌信号可能与 Hp 感染活动期相关。

图 5

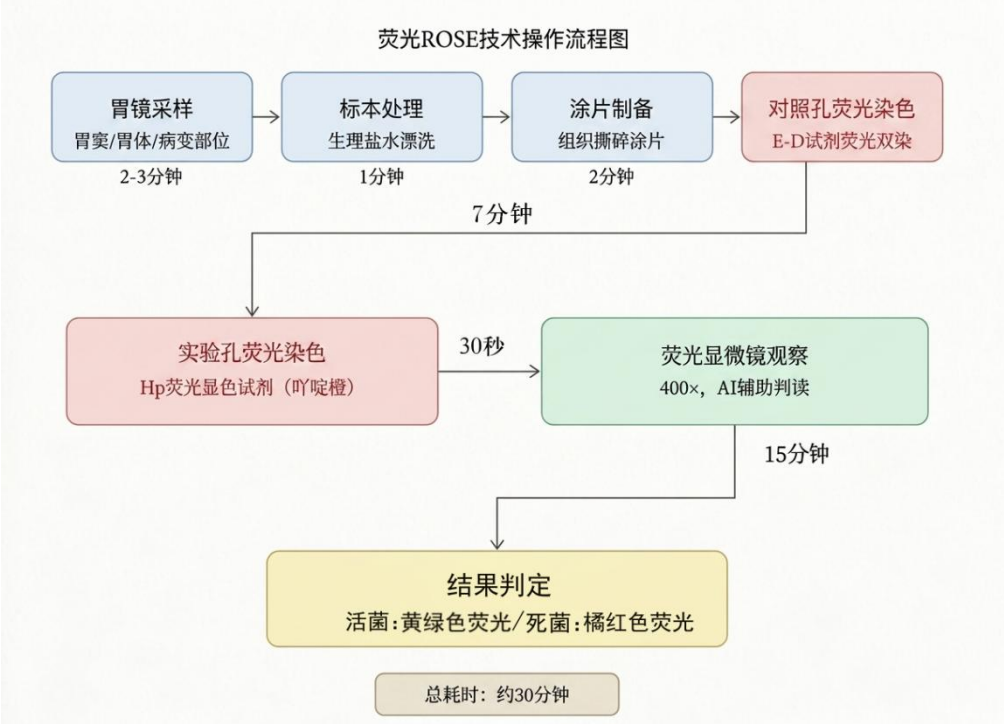
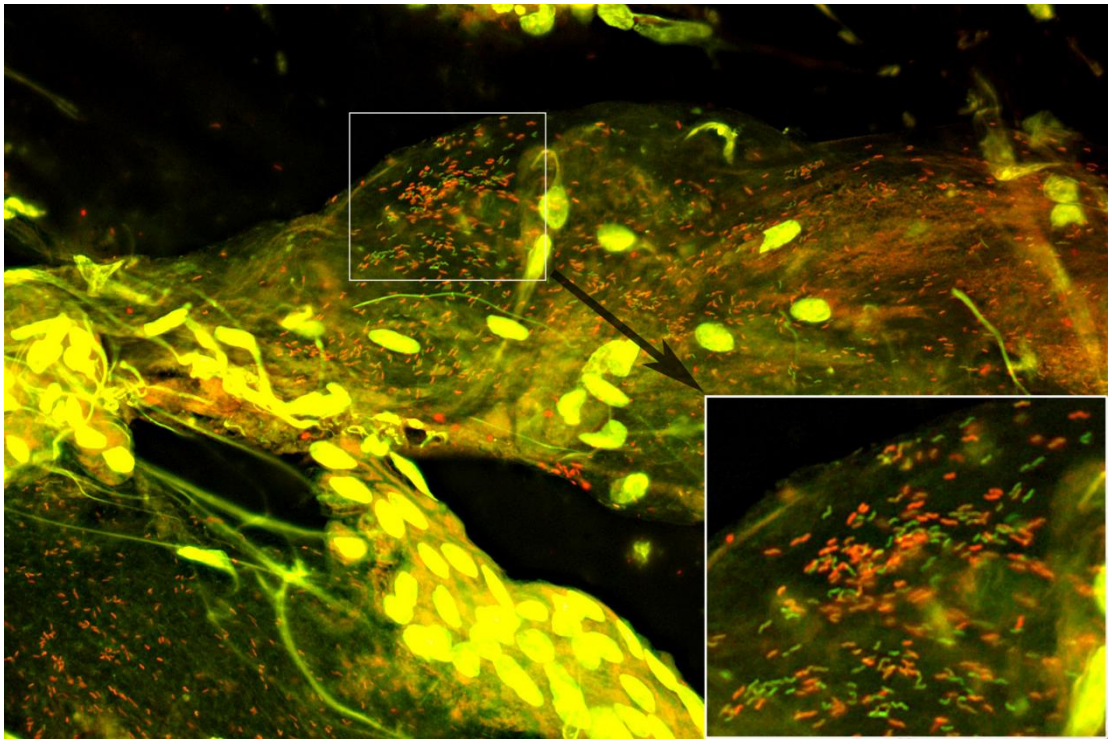


图 6



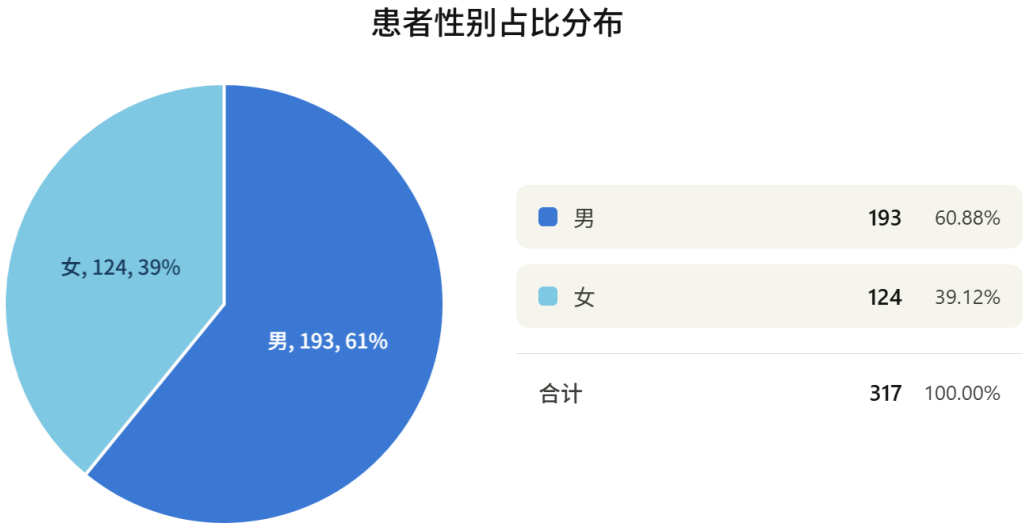
注：Hp 荧光 ROSE 法染色结果（×400 倍），图中发绿色荧光短杆状物为活菌；发橙红色荧光短杆状物为死菌。

3.4 性别与年龄分布

本研究按性别（男 193 例 60.88%、女 124 例 39.12%，图 7）和年龄组（≤40 岁 93 例、41~60 岁 134 例、>60 岁 90 例，表 1、图 8）进行阳性率亚组分析，

以探究人口学因素影响。共行 5 种方法的性别与年龄 χ^2 检验各 1 套（计 10 次独立检验），用 Bonferroni 法校正，校正后检验水准 $\alpha'=0.05/10=0.005$（表 2）。性别方面，病理检查男性阳性率 27.5%、女性 15.3%（原始 P=0.017），经校正后不再显著，其余四种方法均无统计学意义（P>0.05），提示性别非主要混杂因素；年龄方面，五种方法均以 ≤40 岁组阳性率最高，与我国 Hp 感染年轻化趋势一致**错误!不能识别的开关参数。**，原始检验组间差异均显著（P 均<0.05），但经校正后均不再显著。各方法 >60 岁组较 ≤40 岁组阳性率下降幅度为 RUT 14.2、UBT 16.3、肉眼 20.2、病理 16.6 个百分点，而 F-ROSE 仅下降 15.1 个百分点（≤40 岁 49.5%、41~60 岁 32.8%、>60 岁 34.4%），降幅较小，且在三个年龄组中均保持除肉眼外四种客观方法的最高检出率（图 8），提示其在中老年人群仍维持稳健检出能力。

图 7



注：317 例研究样本中男女比例图表

表 1　按性别和年龄分层的阳性率比较（χ^2 检验）

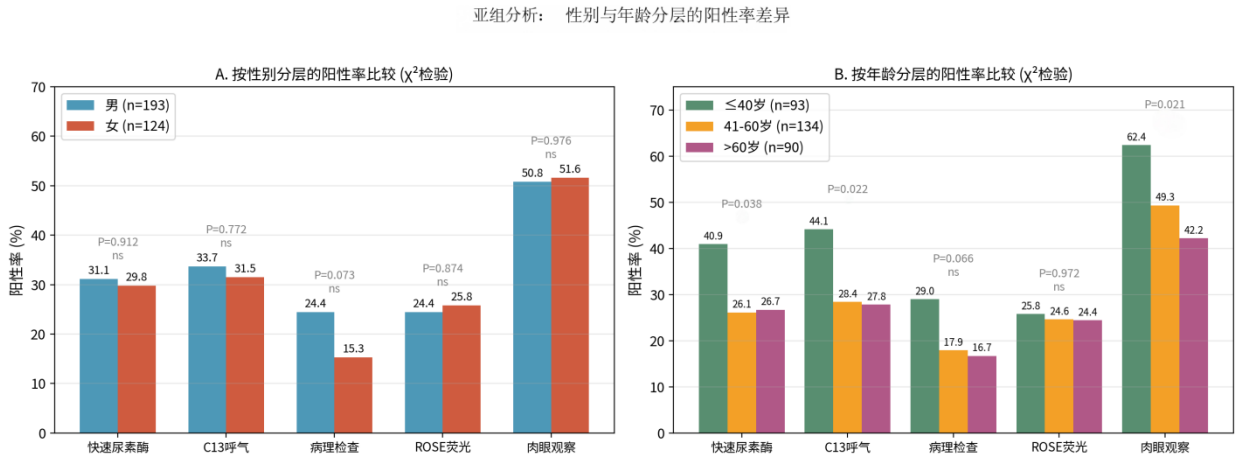
检测方法	男(n=193)	女(n=124)	P(性别)	≤40 岁	41-60 岁	>60 岁	P(年龄)
快速尿素酶	31.1%	29.8%	0.912	40.9%	26.1%	26.7%	0.038
^{13}C-UBT	33.7%	31.5%	0.772	44.1%	28.4%	27.8%	0.022
病理检查	27.5%	15.3%	0.017	33.3%	19.4%	16.7%	0.013
F-ROSE	39.9%	35.5%	0.502	49.5%	32.8%	34.4%	0.028
肉眼观察	50.8%	51.6%	0.976	62.4%	49.3%	42.2%	0.021

注：P<0.05；性别比较（2×2）采用 Pearson χ^2 检验，年龄比较(2×3 列联表)采用 Pearson χ^2 检验。

表 2 Hp 检测方法性别与年龄亚组分析及 Bonferroni 多重比较校正结果

检测方法	男性阳性率(n=193)	女性阳性率(n=124)	性别差异 P 值	≤40 岁阳性率	41~60 岁阳性率	>60 岁阳性率	年龄差异 P 值	Bonferroni 校正后α	校正后显著性
快速尿素酶	31.1%	29.8%	0.912	40.9%	26.1%	26.7%	0.038	0.005	无显著差异
^{13}C-UBT	33.7%	31.5%	0.772	44.1%	28.4%	27.8%	0.022	0.005	无显著差异
病理检查	**27.5%**	**15.3%**	**0.017**	**33.3%**	**19.4%**	**16.7%**	**0.013**	**0.005**	**无显著差异**
F-ROSE	39.9%	35.5%	0.502	49.5%	32.8%	34.4%	0.028	0.005	无显著差异
肉眼观察	50.8%	51.6%	0.976	62.4%	49.3%	42.2%	0.021	0.005	无显著差异

图 8



注：性别分组中除病理检查外（校正前 P=0.017，经校正后均无显著差异），其余方法 P>0.05 标注为 ns，即无显著差异；年龄分组中各方法 P 均<0.05。

3.5 317 例患者五种检测方法阳性率比较

317 例研究对象中五种检测方法的阳性与阴性例数见表 3 及图 9，五种方法阳性率分别为：快速尿素酶试验 30.60%、呼气实验 32.81%、病理观察 22.71%、荧光 ROSE 38.17%、胃镜白光直视判断 51.10%。其中肉眼判断阳性率最高（162/317）；两两比较显示，肉眼观察阳性率显著高于其他 4 种方法（校正 P 均 ＜0.005），病理检查阳性率显著低于其他 4 种方法（校正 P 均 ＜0.005）。为进一步明确各检测方法间的配对结果差异与一致性，本研究采用连续性校正版 McNemar 检验对 5 种方法的二分类阳性 / 阴性结果进行两两配对分析，共完成 10 组对比；同时采用 Bonferroni 法进行多重比较校正，原检验水准 α =0.05，校正后 α ’=0.05/10=0.005，详细统计结果见表 4。结果显示：RUT 与 ^{13}C-UBT、^{13}C-UBT 与 F-ROSE 校正 P 值均远大于 0.005，差异无统计学意义；RUT 与 F-ROSE 校正 P=0.005256，处于临界区间，按校正检验水准亦判定为差异无统计学意义。综合 McNemar 检验与 3.6 节 Kappa 一致性结果（K=0.6943，较强一致），提示三组方法具备较好的检测一致性。其余 7 组两两比较的校正 P 值均＜0.005，差异有统计学意义，其中病理检查、肉眼观察与其余 4 种方法的差异

均有统计学意义（校正 P 均 < 0.005）。

各方法阳性率差异显著，且多数配对检测结果存在统计学上的系统性偏差，提示各方法均存在一定的假阳性或假阴性，无法确定单一可靠的传统金标准，故本研究转而采用 LCA 等无金标准框架进行诊断效能评价。

表 3　五种检测方法在 317 例患者中的阳性率比较

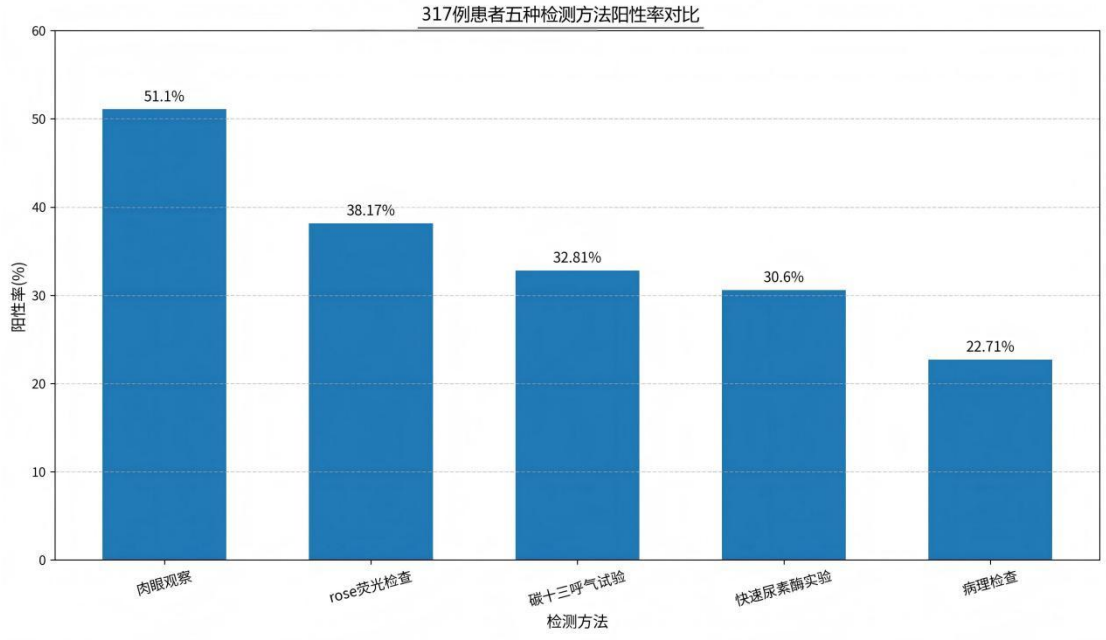
检测方法	总例数	阴性例数	阳性例数	阳性率
肉眼观察	317	155	162	51.10%
rose 荧光检查	317	196	121	38.17%
^{13}C-UBT	317	213	104	32.81%
RUT	317	220	97	30.60%
病理检查	317	245	72	22.71%

注：数据源于本次纳入研究的 317 样本

表 4 McNemar 详细结果

方法 1	方法 2	McNemar 卡方值(连续性校正)	原始 P 值	Bonferroni 校正 P 值	校正后α =0.005，是否显著
RUT	^{13}C-UBT	1.0286	0.310494	1	否
RUT	病理	12.8	0.000347	0.003466	是
RUT	F-ROSE	12.0227	0.000526	0.005256	否
RUT	肉眼观察	47.0805	0	0	是
^{13}C-UBT	病理	17.7963	0.000025	0.000246	是
^{13}C-UBT	F-ROSE	5.9535	0.014688	0.146882	否
^{13}C-UBT	肉眼观察	33.8438	0	0	是
病理	F-ROSE	39.0508	0	0	是
病理	肉眼观察	74.7264	0	0	是
F-ROSE	肉眼观察	21.9178	0.000003	0.000028	是

图 9



3.6　317 例患者五种检测方法整体一致性和方法间一致性分析

对 317 例患者五种幽门螺旋杆菌检测方法进行两两 Kappa 分析显示（表 5），RUT 与^{13}C-UBT 一致性最高（K= 0.745），结果高度吻合；rose 荧光检查与快速尿素酶、^{13}C-UBT 之间也达到强一致（K> 0.69）。肉眼观察与其他四种方法的一致性均偏低，说明肉眼观察存在较大的主观偏差，假阳性率高（图 10）。

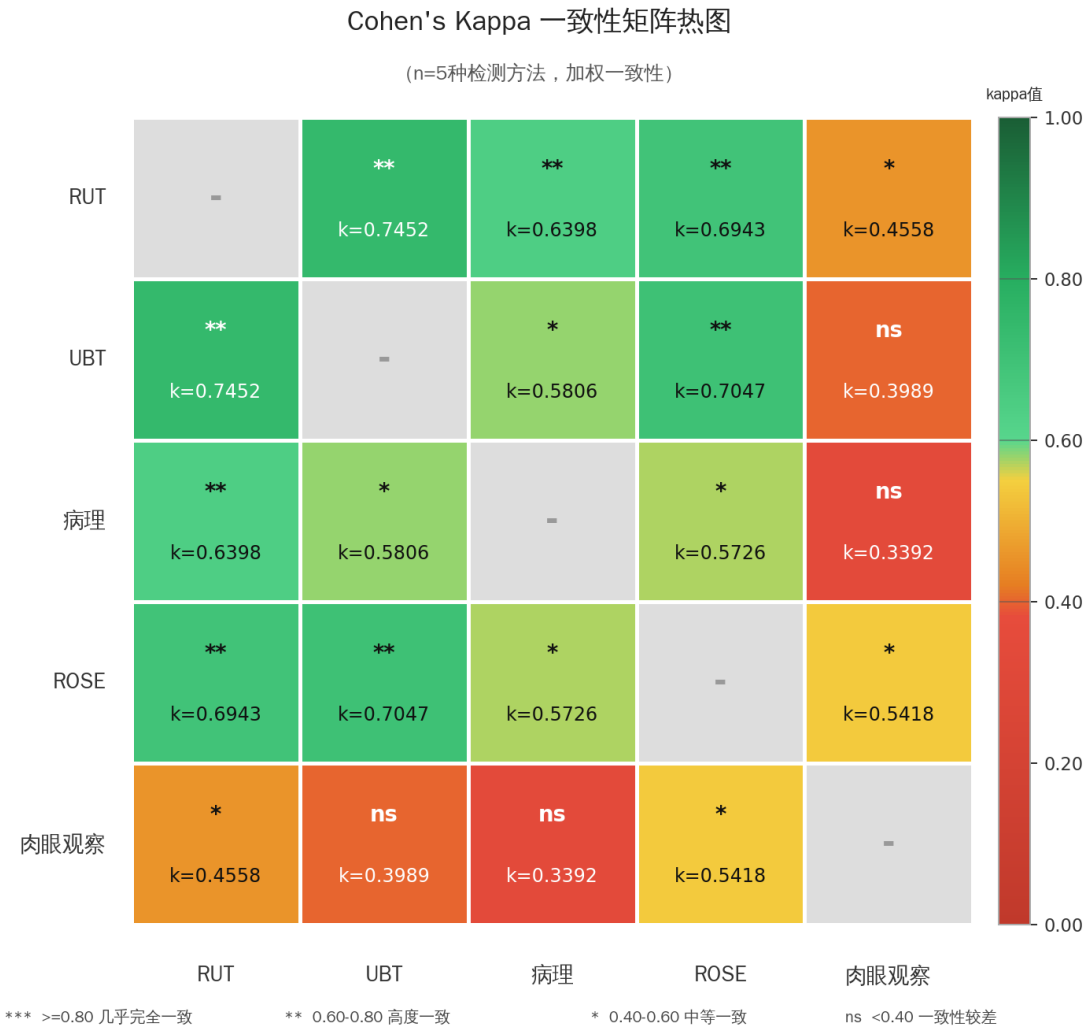
表 5

方法对	观察一致率	期望一致率	Kappa	95%CI	P 值	一致性强度
快速尿素酶测试 vs ^{13}C-UBT 测试	0.8896	0.5667	0.7452	0.6656 - 0.8248	<0.0001	较强
快速尿素酶测试 vs ROSE 荧光检查	0.8612	0.5459	0.6943	0.6105 - 0.7782	<0.0001	较强
快速尿素酶测试 vs 病理检查	0.8580	0.6059	0.6398	0.5423 - 0.7373	<0.0001	较强
^{13}C-UBT 测试 vs ROSE 荧光检查	0.8644	0.5407	0.7047	0.6226 - 0.7867	<0.0001	较强
ROSE 荧光检查 vs 肉眼观察	0.7697	0.4974	0.5418	0.4496 - 0.6340	<0.0001	中等
^{13}C-UBT 测试 vs 病理检查	0.8297	0.5938	0.5806	0.4787 - 0.6825	<0.0001	中等
病理检查 vs ROSE 荧光检查	0.8139	0.5646	0.5726	0.4742 - 0.6710	<0.0001	中等
快速尿素酶测试 vs 肉眼观察	0.7256	0.4957	0.4558	0.3584 - 0.5532	<0.0001	中等
^{13}C-UBT 测试 vs 肉眼观察	0.6972	0.4962	0.3989	0.2985 - 0.4993	<0.0001	一般
病理检查 vs 肉眼观察	0.6656	0.4940	0.3392	0.2366 - 0.4418	<0.0001	一般

注：Kappa 一致性强度判定标准（Landis & Koch，1977）：< 0.20 差，0.20 - 0.40 一般，0.40 - 0.60 中等，0.60 - 0.80 较强，> 0.80 几乎完全一致。

图 10

10.12201/bmr.202606.00055V1



注：317 例患者五种幽门螺旋杆菌检测方法一致性 Kappa 系数热力图

3.7 潜在类别分析（LCA）模型选择与结果分析

3.7.1 LCA 模型选择

本研究分别拟合 K=1、2、3、4 的潜在类别模型，结果见表 6。K=1→2 与 K=2→3 的 Bootstrap 似然比检验（BLRT）均具有统计学意义（均 P<0.001），提示增加类别数可显著改善模型拟合；而 K=3→4 的 BLRT 无统计学意义（P=0.080），提示 K=4 模型存在过拟合。从信息准则来看，AIC、BIC、aBIC 均在 K=3 时最小（BIC=1415.25，较 K=2 的 1432.61 低 17.36），K=3 的熵值（Entropy=0.949）也略高于 K=2（Entropy=0.920）。按纯统计标准，K=3 为相对最优模型，DELTA-BIC=17.36 在通行标准（>10 为强证据）下属较强统计证据，对此本文予以正视，并进一步分析 K=3 中间类的实质内涵。

但本研究最终选择 K=2 模型，理由如下：

1、研究估计目标要求二分类潜变量（统计层面）：本研究的核心目标是估计 5 种检测方法相对于真实感染状态的敏感性（Se）与特异性（Sp）。Se/Sp 在数学定义上以二分类潜在状态（感染 vs 未感染）为前提：Se = P（阳性 | 感染类），Sp = P（阴性 | 未感染类）。K=3 引入第三个中间类后，Se/Sp 的计算依赖于如何定义感染与未感染的边界（中间类并入感染侧还是未感染侧），使核心估计量产生不确定性，无法给出研究问题所需的唯一确定答案。此外，临床上幽门螺旋杆菌感染的诊断与治疗均为二分类决策，国内外指南均无中间感染状态的定义与处理方案，K=3 中间类在临床上无法转化为明确决策。因此 K=2 与研究问题的结构最为匹配。

2、K=3 中间类更可能反映局部独立性假设的轻微违反，而非真实临床亚类：K=3 的 Class 2 为中间类（估计比例 15.1%，模态归类 44 例），其概率剖面具有高度特征性：F-ROSE 阳性率 100%，肉眼观察阳性率 77.7%，而 RUT、^{13}C-UBT 及病理阳性率均低于 40%（表 7）。K=3 的未感染类（Class 1）F-ROSE 阳性率为 0，提示 F-ROSE 对分类起主导作用。结合临床操作流程，内镜医师发现肉眼疑似阳性病灶时常会告知同台 F-ROSE 取样操作者，两者间存在信息关联，这正是局部独立性假设的典型违反场景。因此，最简洁的解释是：K=3 新增的第三类在很大程度上是在吸收肉眼观察与 F-ROSE 之间的方法学相关性，反映了局部独立性假设的轻微违反，而非独立存在的临床感染状态**错误！不能识别的开关参数。**。尽管不能完全排除该类为早期或潜在感染的可能，但无直接临床证据支持。

3、K=2 绝对拟合优度满足，无统计必要升至 K=3：绝对拟合优度检验显示 K=2 模型无显著失拟（P > 0.05），说明模型在整体水平上可充分拟合数据。肉眼观察与 F-ROSE 之间的轻微相关性属于局部、有限的假设偏离，不足以推翻模型整体有效性。同时 K=2 分类清晰度高（Entropy=0.920），BIC 降低仅反映相对拟合优化，不构成升级至 K=3 的充分理由。

4、K=2 与 K=3 对主体人群分类高度一致：两种模型对 191 例明确未感染、82 例明确感染的患者分类完全一致，合计占 86%（表 8）。K=3 仅拆分出 K=2 中约 14% 的边界病例，并未改变核心分类结果。

5、诊断效能评价结果高度稳健：五种检测方法的敏感性、特异性及 Youden 指

数排序在 K=2 与 K=3（方案 A 与方案 B）之间高度一致（表 9），提示无论选择 K=2 或 K=3，核心诊断结论不变。在统计改善有限、临床解释性不足的情况下，选择参数更少、解释更清晰、与研究问题结构匹配的 K=2 模型符合模型简约性原则。

表 6　LCA 模型选择（K=1~4）

K	参数数	对数似然	AIC	BIC	aBIC	Entropy	LR（vs K−1）	BLRT p
1	5	−996.09	2002.18	2020.98	2005.12	—	—	—
2	11	−684.63	1391.26	1432.61	1397.72	0.920	622.92	<0.001
3	17	−658.67	1351.35	1415.25	1361.33	0.949	51.92	<0.001
4	23	−652.61	1351.22	1437.67	1364.72	0.883	12.13	0.080

注：K：潜在类别数；AIC：赤池信息准则；BIC：贝叶斯信息准则；aBIC：校正贝叶斯信息准则；Entropy：熵值（越接近 1 分类越清晰）；LR（vs K-1）：似然比；BLRT p：Bootstrap 似然比检验 p 值（基于 1000 次参数自助）。AIC、BIC、aBIC 值越小表示相对拟合效果越好。

表 7　K=2 与 K=3 模型的类别患病率及条件响应概率

	K = 2 模型	K = 2 模型	K = 3 模型	K = 3 模型	K = 3 模型
检测方法	C1 未感染	C2 感染	C1 未感染	C2 中间类	C3 感染
类别患病率 π	0.6520	0.3480	0.6020	0.1510	0.2470
RUT	0.0253	0.8322	0.0247	0.2913	1.0000
^{13}C-UBT	0.0536	0.8425	0.0405	0.3897	0.9906
病理	0.0157	0.6234	0.0157	0.2006	0.7581
F-ROSE	0.0800	0.9472	0.0000	1.0000	0.9324
肉眼	0.3019	0.9031	0.2732	0.7769	0.9272

注：K=2、K=3 的类别均按"阳性总概率升序"排序，故"C1 未感染"在两个模型中含义可比。

表 8　K=2 模态分类 × K=3 模态分类交叉表（n = 317）

	K=3 C1 未感染	K=3 C2 中间类	K=3 C3 感染	合计
K=2 C1 未感染	191	20	0	**211**
K=2 C2 感染	0	24	82	**106**
合计	**191**	**44**	**82**	**317**

注：K=3 的中间类（n=44）由 K=2 中分类最不确定的边界患者构成——20 例原被 K=2 归为未感染、24 例原被归为感染。

两个模型在置信度极高的两端（191 例确诊未感染 / 82 例确诊感染）完全一致，K=3 的真实作用是识别出约 14% 的边界患者，而非根本性改变多数患者的归类。

表 9　五种检测方法在 K=2 与 K=3 不同定义下的 Se 与 Sp（稳健性分析）

	K=2	K=2	K=3 方案 A（仅 C3 为感染）		K=3 方案 B（C2+C3 合并为感染）	
检测方法	Se	Sp	Se	Sp	Se	Sp
人群感染率	34.79%	34.79%	24.71%	24.71%	39.84%	39.84%
RUT	0.8322（0.7397, 0.9151）	0.9747（0.9431, 1.0000）	1.0000（0.97, 1.00）	0.9217（0.85, 0.97）	0.7308（0.65, 0.80）	0.9754（0.95, 1.00）
^{13}C-UBT	0.8425（0.7489, 0.9249）	0.9464（0.9010, 0.9815）	0.9906（0.93, 1.00）	0.8893（0.83, 0.93）	0.7623（0.66, 0.83）	0.9595（0.94, 0.99）
病理	0.6234（0.4998, 0.7449）	0.9843（0.9624, 1.0000）	0.7581（0.64, 0.94）	0.9471（0.92, 0.98）	0.5463（0.43, 0.62）	0.9843（0.96, 1.00）
F-ROSE	0.9472（0.8763, 0.9975）	0.9200（0.8700, 0.9612）	0.9324（0.87, 1.00）	0.7990（0.74, 0.86）	0.9580（0.84, 0.98）	1.0000（0.99, 1.00）
肉眼	0.9031（0.8249, 0.9682）	0.6981（0.6101, 0.7820）	0.9272（0.86, 0.98）	0.6255（0.57, 0.69）	0.8701（0.79, 0.92）	0.7268（0.67, 0.80）

注：K=3 方案 A：仅以 C3（感染类）为感染，C1+C2 合并为未感染；方案 B：C2+C3 合并为感染，C1 为未感染。两种方案下排序与 K=2 高度一致，诊断效能结论稳健。

3.7.2 潜在类别分析（LCA）结果分析

采用 2 类别潜在类别分析模型（EM 算法，60 次随机起始点）对 5 种检测方法的联合反应模式进行建模。最终模型对数似然值为−684.63，模型估计的人群潜在 Hp 感染患病率为 34.79%（Bootstrap 95%CI 29.3%~41.7%）。各检测方法的诊断性能如表 10 所示。在敏感性方面，ROSE 荧光检查最高（Se=0.9472，95%CI 0.8763~0.9975），肉眼观察次之（Se=0.9031，95%CI 0.8249~0.9682），病理检查最低（Se=0.6234，95%CI 0.4998~0.7449）。在特异性方面，病理检查最高（Sp=0.9843，95%CI 0.9624~1.0000），RUT 次之（Sp=0.9747，95%CI 0.9431~1.0000），肉眼观察最低（Sp=0.6981，95%CI 0.6101~0.7820）。

敏感性与特异性森林图（图 11）显示，F-ROSE 敏感性 0.947 的置信区间上限接近 1.0，特异性的置信区间宽度明显窄于敏感性，提示其特异性估计更稳定。

以约登指数（Youden index, J）衡量综合诊断性能，5 种方法排序（图 12）为：ROSE 荧光（J=0.8672）> RUT（J=0.8069）>^{13}C-UBT（J=0.7889）> 病理检查

（J=0.6076）≈ 肉眼观察（J=0.6013）。ROSE 荧光检查的综合判别能力最优，其阴性预测值亦最高（NPV=0.9703，95%CI 0.9380~0.9975），对于排除 Hp 感染具有较高价值。肉眼观察虽敏感性较高，但特异性不足，阳性预测值仅 0.6148（95%CI 0.5340~0.6937），独立用于确诊时需谨慎。

表 10　LCA 模型估计的五种检测方法诊断性能指标

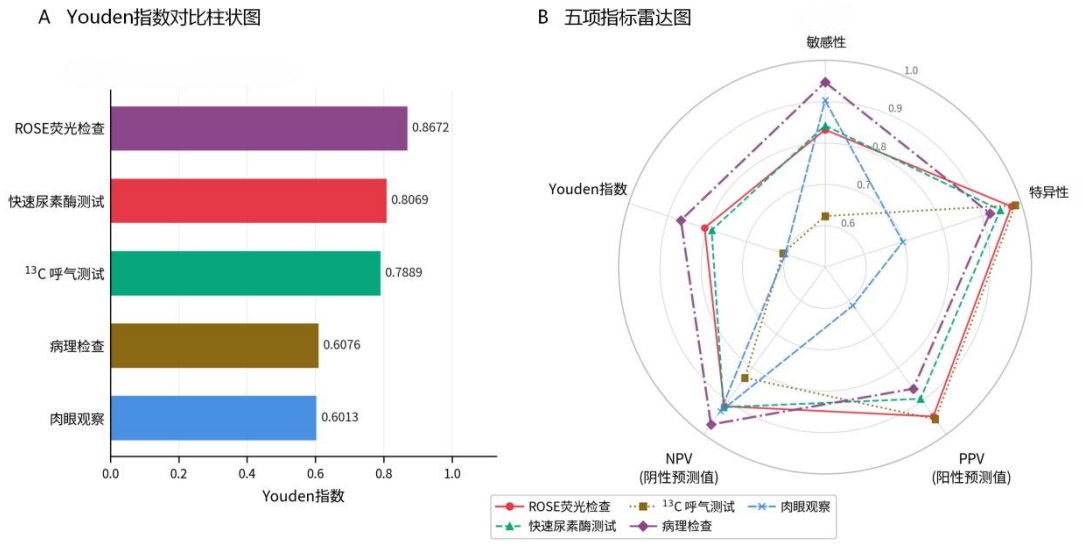
检测方法	敏感性	95%CI	特异性	95%CI	阳性预测值	95%CI	阴性预测值	95%CI	Youden 指数	95%CI
快速尿素酶测试	0.8322	[0.7397, 0.9151]	0.9747	[0.9431, 1.0000]	0.9461	[0.8932, 0.9924]	0.9159	[0.8696, 0.9576]	0.8069	[0.6950, 0.9108]
^{13}C-UBT 测试	0.8425	[0.7489, 0.9249]	0.9464	[0.9010, 0.9815]	0.8934	[0.8205, 0.9535]	0.9185	[0.8752, 0.9557]	0.7889	[0.6682, 0.8966]
病理检查	0.6234	[0.4998, 0.7449]	0.9843	[0.9624, 1.0000]	0.9548	[0.8959, 0.9973]	0.8305	[0.7753, 0.8815]	0.6076	[0.4740, 0.7367]
ROSE 荧光检查	0.9472	[0.8763, 0.9975]	0.9200	[0.8700, 0.9612]	0.8633	[0.7926, 0.9219]	0.9703	[0.9380, 0.9975]	0.8672	[0.7715, 0.9472]
肉眼观察	0.9031	[0.8249, 0.9682]	0.6981	[0.6101, 0.7820]	0.6148	[0.5340, 0.6937]	0.9311	[0.8794, 0.9756]	0.6013	[0.4780, 0.7283]

注：Youden 指数 = 敏感性 + 特异性 − 1；95%CI 由 Bootstrap 500 次计算得出。

图 11

注：每条数据由三个视觉元素组成：实心圆点（点估计值）、横线（95% 置信区间范围）、竖向端帽（CI 的上下边界）。CI 越宽代表不确定性越大，通常意味着该研究的样本量较小或数据离散度高。

图 12



注：A：Youden 指数=敏感性+特异性-1，数值越高越好；B：外圈 = 1.0：也就是每个指标的 “满分”，越靠近外圈，代表这个指标的数值越高、表现越好。

3.7.3 贝叶斯敏感性分析结果

采用贝叶斯潜在类别模型（BLCM）对 EM-MLE 结果进行敏感性验证。9 个文献来源参数采用信息性 Beta 先验，F-ROSE 的 Se、Sp 采用 Beta (1,1) 非信息先验（表 11），所有信息性先验等效样本量上限设为 50，确保以本研究数据为主导后验估计。Gibbs 抽样 3 条独立链，经收敛诊断（$\hat{R}$均 < 1.05，有效样本量充足），提示模型收敛良好。

贝叶斯主分析（方案 A）后验估计与 EM-MLE 结果对比如下（表 12）：潜在患病率π 的贝叶斯后验中位数为 0.332（95% HPDI 0.280 – 0.386），与 EM-MLE 的 0.3479 高度一致。F-ROSE 的 Se 后验中位数 0.948（95% HPDI 0.889 – 0.983）与 EM-MLE 的 0.947 几乎完全重合（差异 0.1%），Sp 后验中位数 0.891（95% HPDI 0.838 – 0.936）与 EM-MLE 的 0.920 相差 2.9 个百分点，在合理统计波动范围内，证实 F-ROSE 诊断性能估计稳健。RUT、^{13}C-UBT、肉眼观察的后验估计与 EM-MLE 差异均较小（<5%）；病理检查 Se 的贝叶斯后验中位数 0.749 较 EM-MLE 的 0.623 上调 12.6 个百分点，反映文献 pooled Se=93% 对本中心数据的借力修正效应，提示本中心病理操作流程可能存在改进空间，但此修正未改变病理在诊断效能排序中的位置。

三种先验方案对比（表 12）显示：按 “差异≤5% 判定稳健” 的标准，除病理 Se 外，其余方法的 Se、Sp 在三种先验下点估计差异均 <5%。其中 F-ROSE 的 Se、

Sp 最大差异分别仅为 0.97%、3.02%，稳定性最优；病理 Se 三方案点估计为 0.749/0.616/0.803，最大差异达 18.7%，对先验选择敏感。后验分布显示（图 13A－C），幽门螺旋杆菌潜在患病率、F-ROSE 敏感性、F-ROSE 特异性在方案 A（主分析）、方案 B（全非信息）、方案 C（强先验）下高度集中，中位数接近、分布重叠度高，提示 F-ROSE 及患病率估计高度稳定。五种方法敏感性与特异性三方案对比进一步验证（图 13D、E）：F-ROSE 表现最稳定，其余方法除病理 Se 外均满足稳健标准。LCA 与 BLCM 患者后验感染概率对比显示两模型概率高度相关，沿无差异线集中，硬分类一致率达 95.0%；差值直方图呈对称分布，均值≈0，提示 LCA 主结果整体稳健可靠（图 13F、G）。F-ROSE 的 Sp 在贝叶斯后验中较 EM-MLE 下调 2.9 个百分点，此差异源于 Beta(1,1)非信息先验对极端值（Sp 接近 1.0）的收缩效应，属贝叶斯估计的合理正则化表现，不改变 F-ROSE 综合效能最优的核心结论。

表 11 先验分布参数设定

参数	文献 pooledmean	95% CI	Beta 分布α	Beta 分布β	ESS(α +β)
潜在患病率 π	0.379	0.350 - 0.408	18.95	31.05	50
Se_RUT	0.940	0.870 - 0.970	47.00	3.00	50
Sp_RUT	0.910	0.790 - 0.960	38.72	3.83	42.5
Se_^{13}C-UBT	0.791	0.742 - 0.825	39.55	10.45	50
Sp_^{13}C-UBT	0.750	0.700 - 0.786	37.50	12.50	50
Se_病理检查	0.930	0.880 - 0.960	46.50	3.50	50
Sp_病理检查	0.840	0.660 - 0.930	22.96	4.37	27.3
Se_肉眼观察	0.816	0.790 - 0.841	40.80	9.20	50
Sp_肉眼观察	0.868	0.850 - 0.884	43.40	6.60	50
Se_F-ROSE	—	—	1.00	1.00	2（非信息性）
Sp_F-ROSE	—	—	1.00	1.00	2（非信息性）

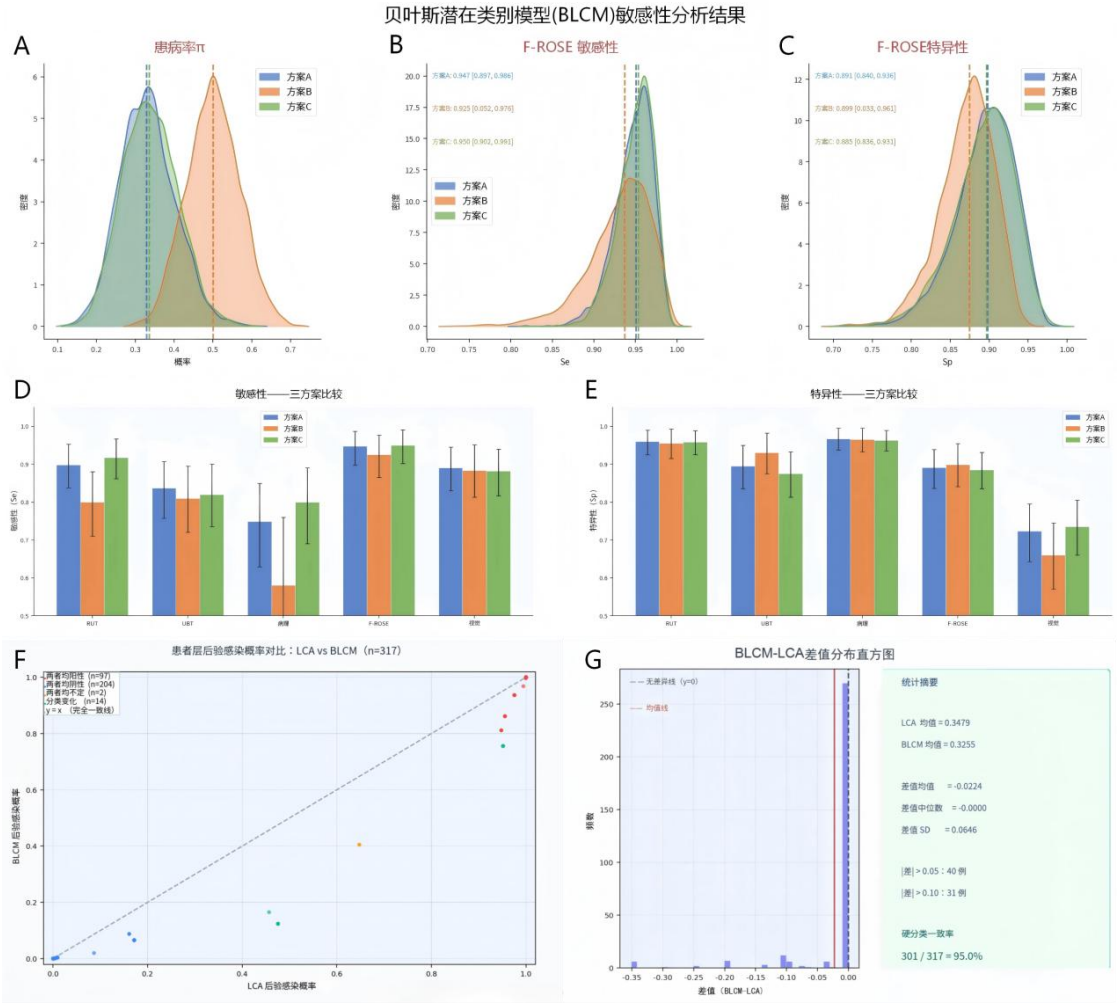
注：所有信息性先验的等效样本量 ESS 设上限为 50，相对于本研究 317 例样本量保证数据主导后验估计；F-ROSE 采用 Beta(1,1)非信息性先验。

表 12　贝叶斯敏感性分析下五种方法 Se/Sp 与 EM-MLE 对比

参数	方案 A 主分析	方案 B 全非信息性	方案 C 强先验	EM-MLE
π	0.332 (0.280 - 0.386)	0.349 (0.292 - 0.415)	0.334 (0.287 - 0.381)	0.348
Se_RUT	0.899 (0.836 - 0.947)	0.826 (0.735 - 0.904)	0.911 (0.860 - 0.952)	0.832
Sp_RUT	0.959 (0.929 - 0.981)	0.972 (0.939 - 0.991)	0.959 (0.927 - 0.980)	0.975
Se_UBT	0.837 (0.769 - 0.893)	0.837 (0.748 - 0.906)	0.825 (0.769 - 0.873)	0.843
Sp_UBT	0.895 (0.852 - 0.931)	0.944 (0.902 - 0.973)	0.868 (0.825 - 0.904)	0.946
Se_病理	0.749 (0.676 - 0.819)	0.616 (0.520 - 0.713)	0.803 (0.741 - 0.859)	0.623
Sp_病理	0.966 (0.936 - 0.986)	0.981 (0.954 - 0.994)	0.968 (0.939 - 0.985)	0.984
Se_F-ROSE	**0.948 (0.889 - 0.983)**	0.941 (0.885 - 0.978)	0.951 (0.895 - 0.985)	**0.947**
Sp_F-ROSE	**0.891 (0.838 - 0.936)**	0.916 (0.862 - 0.964)	0.886 (0.833 - 0.932)	**0.920**
Se_肉眼	0.891 (0.834 - 0.936)	0.900 (0.825 - 0.954)	0.872 (0.823 - 0.914)	0.903
Sp_肉眼	0.722 (0.664 - 0.777)	0.696 (0.633 - 0.759)	0.743 (0.691 - 0.792)	0.698

注：方案 A 为主分析（9 个文献先验 ESS 截断为 50 + F-ROSE Beta(1,1)非信息性先验）；方案 B 为全非信息性先验 Beta(1,1)；方案 C 为强先验（信息性先验 ESS 截断为 100）。括号内为 95% HPDI。粗体为 F-ROSE 参数及其 EM-MLE 对照。

图 13



注：A－C：分别为幽门螺旋杆菌潜在患病率、F-ROSE 敏感性、F-ROSE 特异性在三种先验方案下的后验分布；蓝、红、绿分别代表方案 A、B、C，竖线为后验中位数。D－E：五种方法敏感性与特异性的三方案对比，柱高为后验中位数，误差棒为 95% HPDI。F：LCA 与 BLCM 个体后验感染概率相关性散点图。G：两模型后验概率差值分布直方图。

3.8 各检测方法 ROC 曲线结果及解读

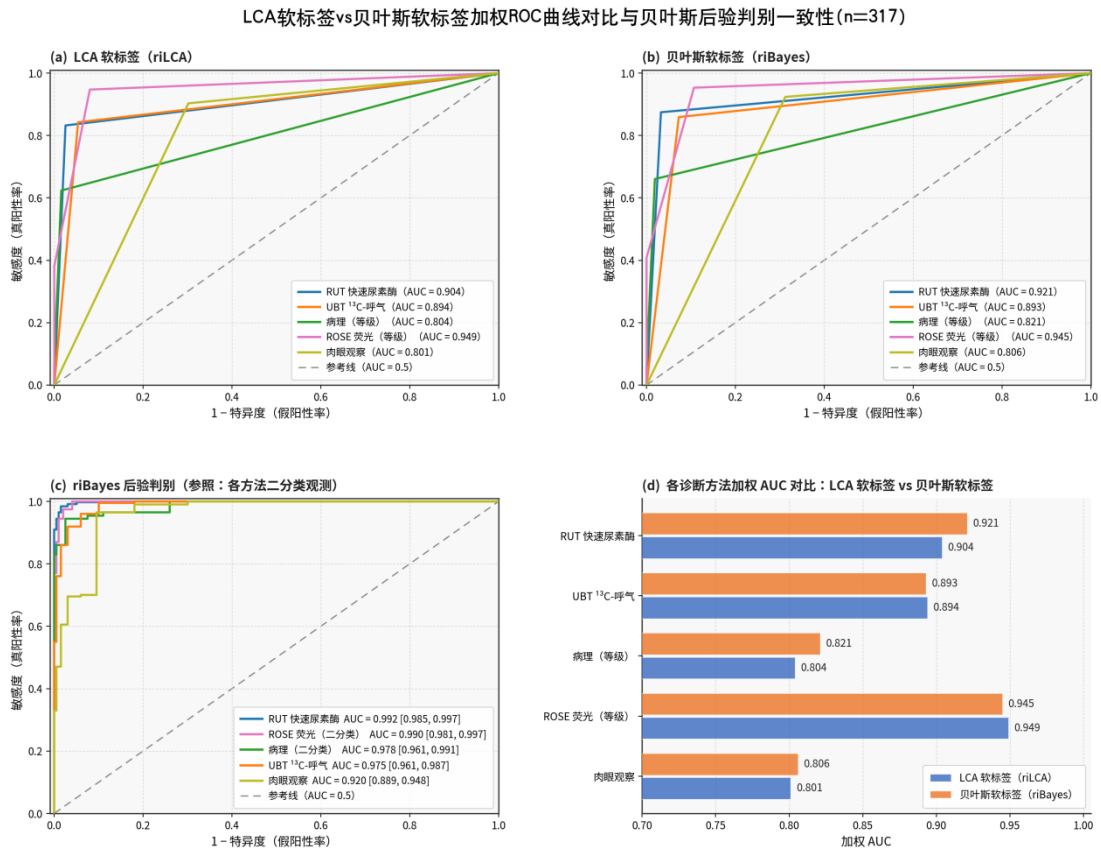
为评估贝叶斯模型估计的共识概率与各检测方法实测结果的判别一致性，本研究在同一队列（n = 317）中分别以 LCA 软标签（r_i^{LCA}）和贝叶斯软标签（r_i^{Bayes}）作为参照，对五种 Hp 检测方法绘制了加权 ROC 曲线，并进一步以 r_i^{Bayes} 后验概率为分级、各方法二分类阳/阴性观测结果为参照绘制了一致性 ROC 曲线。在以 r_i^{LCA}为软标签的加权 ROC 分析中（图 14a），ROSE 荧光（等级）的判别效能最高（AUC = 0.949），其后依次为 RUT 快速尿素酶（AUC = 0.904）、UBT ^{13}C-呼气试验（AUC = 0.894）、病理等级（AUC = 0.804）与胃镜白光直视观察（AUC = 0.801）。改以 r_i^{Bayes} 为软标签后（图 14b），各方法 AUC 排序保持

不变（ROSE 荧光 0.945 > RUT 0.921 > UBT 0.893 > 病理 0.821 > 肉眼 0.806），RUT 与病理在贝叶斯框架下分别提升约 0.017 和 0.017，其余三种方法的 AUC 变化幅度均 ≤ 0.005。两套框架下加权 AUC 的横向对比见图 14d——五种方法两套软标签的 AUC 绝对差异均不超过 0.02，提示贝叶斯共识概率与 LCA 软标签在五种方法上给出的判别效能高度可比，模型选择对该结论稳健。

以r_i^{Bayes} 后验概率为评分、各方法独立二分类结果为参照的一致性 ROC 曲线显示（图 14c）：贝叶斯模型估计的共识概率与各方法实测判定具有很高的吻合度，RUT（AUC = 0.992，95% CI：0.985–0.997）、F-ROSE（AUC = 0.990，95% CI：0.981–0.997）、病理检查（AUC = 0.978，95% CI：0.961–0.991）、^{13}C-呼气试验（AUC = 0.975，95% CI：0.961–0.987）的 AUC 均≥0.975，胃镜白光直视观察为 0.920（95% CI：0.889–0.948）。该结果仅反映贝叶斯后验概率与各单项检测结果的内部一致性，因后验概率由五种方法联合构建，不宜作为独立外部验证证据。除肉眼观察外，其余四种方法的 AUC 区间高度重叠，提示 r_i^{Bayes} 与各单项方法判定均保持稳定一致性，可作为综合反映 Hp 感染概率的合理代理指标。

上述 ROC 结果与 3.7.2 节 LCA、3.7.3 节贝叶斯敏感性分析所得灵敏度、特异度及 Youden 指数结论完全吻合，从多条独立分析路径共同支持 F-ROSE 为本研究中诊断效能最优的 Hp 检测方法这一核心结论。

图 14



注：(a) 以 LCA 软标签为参照的加权 ROC 曲线；(b) 以贝叶斯软标签为参照的加权 ROC 曲线；(c) 以贝叶斯后验概率为评分、各方法二分类结果为参照的一致性 ROC 曲线；(d) 两种软标签框架下 AUC 横向对比。AUC 采用梯形法计算，95% CI 由 Bootstrap 法（B=1000）估计。对角虚线为无判别参考线（AUC=0.5）。

3.9 DeLong 检验

采用加权 DeLong 法对两种软标签金标准（riLCA、riBayes）下五种诊断方法的加权 ROC 曲线进行两两比较，以检验 AUC 差异的统计学意义，检验水准 α=0.05。结果显示，两种软标签框架下方法间效能差异的显著性格局总体一致。在 riLCA 软标签框架下，ROSE 荧光的 AUC 最高（0.9488），显著高于 RUT（P=0.0456）、UBT（P=0.0162）、病理（P<0.001）及肉眼观察（P<0.001）。RUT 与 UBT 之间（P=0.6879）和病理与肉眼观察之间（P=0.9069）差异无统计学意义。

在 riBayes 软标签框架下，方法 AUC 排序与 riLCA 一致，ROSE 荧光（0.9450）仍为最高。ROSE 荧光显著高于 UBT（P=0.0212）、病理（P<0.001）及肉眼观察（P<0.001），但与 RUT 的差异无统计学意义（P=0.2654）。两两比较结果提示，ROSE 荧光诊断效能显著优于病理、肉眼观察及 UBT，与 RUT 效能相当。两套软标签得到的方法效能排序一致；ROSE 显著优于病理、肉眼观察及 UBT，与 RUT 的差异在两套参照下临界（LCA P=0.0456，贝叶斯 P=0.2654），表明本研究结果对金标准定义具有良好稳健性（表 13，图 15，图 16）。

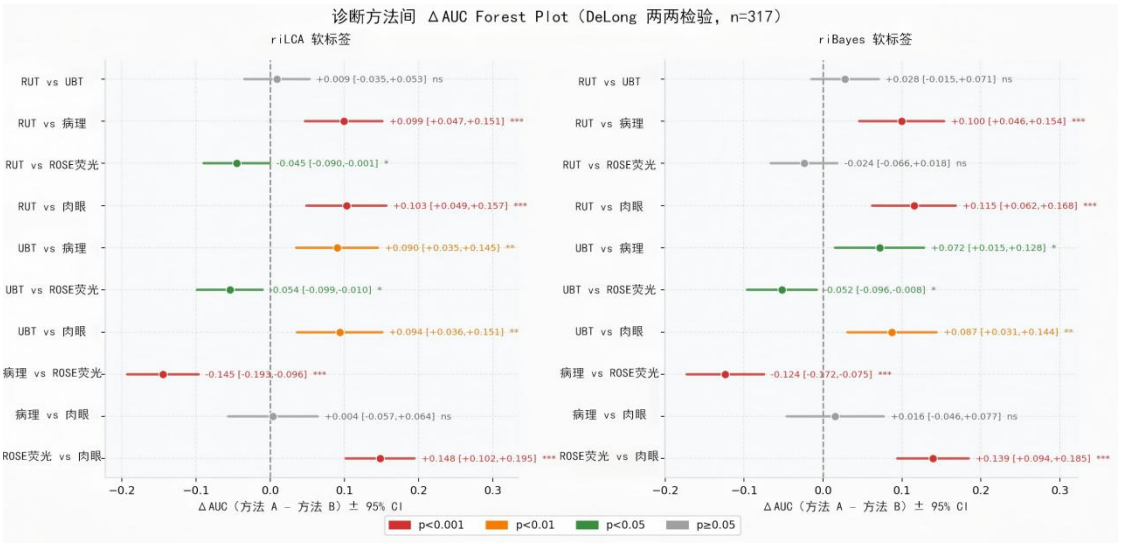
表 13

各诊断方法两两加权 DeLong 检验结果（n=317）

面板	方法 A	AUC(A)	SE(A)	方法 B	AUC(B)	SE(B)	Δ AUC (A−B)	95%CI 下限	95%CI 上限	Z 统计量	p 值	显著性
riLCA 软标签	RUT	0.9035	0.0187	UBT	0.8944	0.0191	0.009	−0.035	0.053	0.402	0.6879	ns
riLCA 软标签	RUT	0.9035	0.0187	病理	0.8042	0.0236	0.0992	0.0471	0.1513	3.733	<0.001	***
riLCA 软标签	RUT	0.9035	0.0187	ROSE 荧光	0.9488	0.0126	−0.0454	−0.0899	−0.0009	−1.999	0.0456	*
riLCA 软标签	RUT	0.9035	0.0187	肉眼观察	0.8006	0.0214	0.1028	0.049	0.1567	3.745	<0.001	***
riLCA 软标签	UBT	0.8944	0.0191	病理	0.8042	0.0236	0.0902	0.035	0.1454	3.201	0.0014	**
riLCA 软标签	UBT	0.8944	0.0191	ROSE 荧光	0.9488	0.0126	−0.0544	−0.0987	−0.0101	−2.405	0.0162	*
riLCA 软标签	UBT	0.8944	0.0191	肉眼观察	0.8006	0.0214	0.0938	0.0364	0.1512	3.203	0.0014	**
riLCA 软标签	病理	0.8042	0.0236	ROSE 荧光	0.9488	0.0126	−0.1446	−0.1928	−0.0964	−5.879	<0.001	***
riLCA 软标签	病理	0.8042	0.0236	肉眼观察	0.8006	0.0214	0.0036	−0.0568	0.064	0.117	0.9069	ns
riLCA 软标签	ROSE 荧光	0.9488	0.0126	肉眼观察	0.8006	0.0214	0.1482	0.1016	0.1948	6.231	<0.001	***
riBayes 软标签	RUT	0.9211	0.0175	UBT	0.8932	0.0194	0.0279	−0.0149	0.0706	1.278	0.2011	ns
riBayes 软标签	RUT	0.9211	0.0175	病理	0.8214	0.0239	0.0997	0.0458	0.1536	3.623	<0.001	***
riBayes 软标签	RUT	0.9211	0.0175	ROSE 荧光	0.945	0.0129	−0.024	−0.0662	0.0182	−1.114	0.2654	ns
riBayes 软标签	RUT	0.9211	0.0175	肉眼观察	0.8058	0.0206	0.1153	0.0624	0.1682	4.273	<0.001	***
riBayes 软标签	UBT	0.8932	0.0194	病理	0.8214	0.0239	0.0718	0.0155	0.1282	2.498	0.0125	*
riBayes 软标签	UBT	0.8932	0.0194	ROSE 荧光	0.945	0.0129	−0.0519	−0.096	−0.0078	−2.305	0.0212	*
riBayes 软标签	UBT	0.8932	0.0194	肉眼观察	0.8058	0.0206	0.0874	0.031	0.1438	3.04	0.0024	**
riBayes 软标签	病理	0.8214	0.0239	ROSE 荧光	0.945	0.0129	−0.1237	−0.1724	−0.0749	−4.972	<0.001	***
riBayes 软标签	病理	0.8214	0.0239	肉眼观察	0.8058	0.0206	0.0156	−0.0461	0.0772	0.496	0.6200	ns
riBayes 软标签	ROSE 荧光	0.945	0.0129	肉眼观察	0.8058	0.0206	0.1393	0.0938	0.1847	6.004	<0.001	***

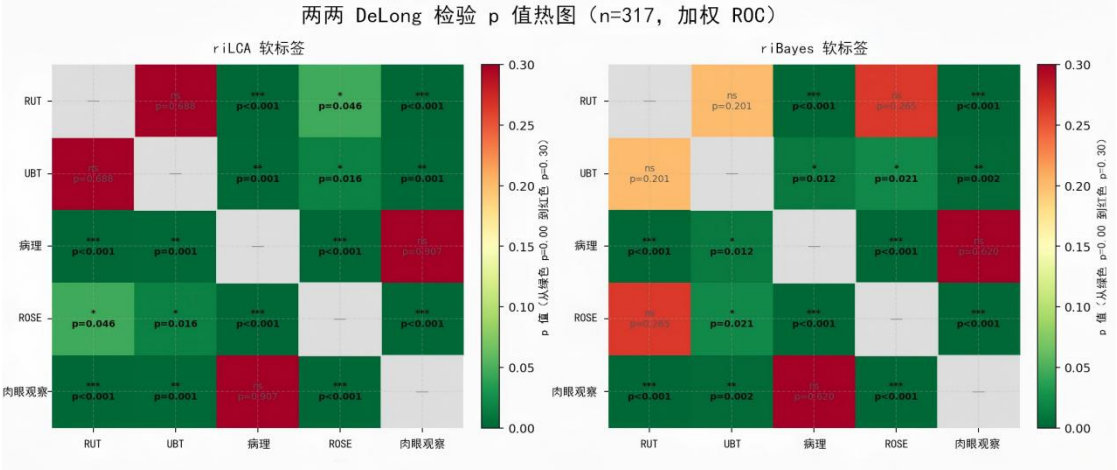
注：Δ AUC=AUC(A)−AUC(B)；*P<0.05，**P<0.01，***P<0.001，ns 为差异无统计学意义；riLCA 与 riBayes 为两套独立软标签，不可跨面板直接比较。

图 15



注：颜色编码：红色 p<0.001，橙色 p<0.01，绿色 p<0.05，灰色 p≥0.05（无统计学意义）。Δ AUC>0 表示方法 A 的 AUC 高于方法 B；Δ AUC<0 表示方法 B 的 AUC 高于方法 A；CI 不跨 0 提示差异具有统计学意义。

图 16



注：颜色由绿色（p=0）到红色（p=0.30）渐变，绿色越深提示 p 值越小、差异越显著；红色越深提示 p 值越大、差异无统计学意义。单元格内标注了精确 p 值及显著性符号（*p<0.05，**p<0.01，***p<0.001，ns 无统计学意义）。

3.10 五种方法两两联合检测策略的诊断效能评价

鉴于五种方法在敏感性、特异性上的互补特征，本研究进一步评估两两联合检测策略的诊断效能。对所有 C(5,2)=10 对方法分别采用并联（OR：任一阳性即视为阳性）与串联（AND：均为阳性才视为阳性）两种规则，共构成 20 个组合策略。各策略的 Se、Sp、PPV、NPV 与 Youden 指数分别基于 LCA 后验概率（软标签加权法）和贝叶斯后验概率（软标签加权法）估计。

1. 各联合方案的整体诊断效能（AUC）

以 LCA 软标签为参照时（表 14），10 种两两联合方案的 AUC 均 ≥ 0.93，整体诊断效能均较高，但相互之间仍存在显著差异。其中以 RUT 联合 F-ROSE 表现最优（AUC = 0.990，95% CI：0.984 – 0.996），其后依次为 ^{13}C-UBT 联合 F-ROSE（0.974，0.959 – 0.989）、RUT 联合肉眼观察（0.972，0.959 – 0.986）、病理联合 F-ROSE（0.967，0.948 – 0.986）、RUT 联合 ^{13}C-UBT（0.962，0.937 – 0.987）、RUT 联合病理（0.960，0.935 – 0.985）、F-ROSE 联合肉眼观察（0.959，0.940 – 0.978）、^{13}C-UBT 联合肉眼观察（0.955，0.931 – 0.978）、^{13}C-UBT 联合病理（0.945，0.916 – 0.973），AUC 最低者为病理联合肉眼观察（0.928，0.901 – 0.955）（图 17A）。

以贝叶斯软标签为参照时（表 15），排序与上述结果高度一致：RUT 联合 F-ROSE 仍居首位（AUC = 0.984，95% CI：0.975 – 0.994），其后依次为 ^{13}C-UBT 联合 F-ROSE（0.974）、RUT 联合肉眼观察（0.971）、^{13}C-UBT 联合肉眼观察（0.967）、

病理联合 F-ROSE（0.966）、RUT 联合 ^{13}C-UBT（0.960）、F-ROSE 联合肉眼观察（0.959）、RUT 联合病理（0.958）、^{13}C-UBT 联合病理（0.954），AUC 最低者仍为病理联合肉眼观察（0.940）(图 17B)。两套参照下结果高度一致，提示 RUT 与 F-ROSE 的联合在 10 种两两组合中具有最佳的整体诊断效能。

2. 串联（AND）与并联（OR）工作点的对照

在 LCA 软标签参照下，串联规则一致表现为"高特异度、低敏感度"：10 种联合方案在串联规则下的特异度均 ≥ 94.3%,但敏感度分布于 57.5%–87.7%。并联规则则表现相反，敏感度多达 94%–100%，其中 RUT+ROSE、UBT+ROSE、RUT+肉眼、ROSE+肉眼、UBT+肉眼等方案敏感度达到或接近 100%（NPV 同时达 100%），但特异度降至 63.5%–96.2%。贝叶斯参照下的工作点参数与 LCA 参照基本一致（典型值差异 < 2 个百分点），结论不变。由此可见，串联规则适用于以"确诊（rule-in）"为目的的临床场景（最大化特异度与 PPV，减少假阳性），并联规则适用于以"筛查或排除诊断（rule-out）"为目的的临床场景（最大化敏感度与 NPV，减少漏诊）。两种规则的选择应基于检测目的而非单纯依据 AUC 大小。

3. 两两 AUC 的 DeLong 比较

配对 DeLong 检验结果显示,RUT 联合 F-ROSE 的 AUC 在 LCA 参照下显著高于其余 9 种联合方案（P 值范围 7.2×10^{-6} 至 0.023，全部 $P < 0.05$）；在贝叶斯参照下亦显著优于 6 种联合方案（除与 ^{13}C-UBT 联合 F-ROSE、^{13}C-UBT 联合肉眼观察、病理联合 F-ROSE 比较 P ≥ 0.05 外，其余两两比较 P < 0.05），进一步证实其整体诊断效能最优。（表 16、17）

4. 小结

综合 AUC 大小、串联/并联工作点的敏感度–特异度组合以及配对 DeLong 检验结果，本研究在两套软标签参照下均得到一致结论：RUT 与 F-ROSE 的两两联合为最优整体诊断方案（AUC 显著最高，并联工作点下 Se = 100%、NPV = 100%）。

表 14 LCA 软标签串联（AND）vs 并联（OR）工作特性对照表

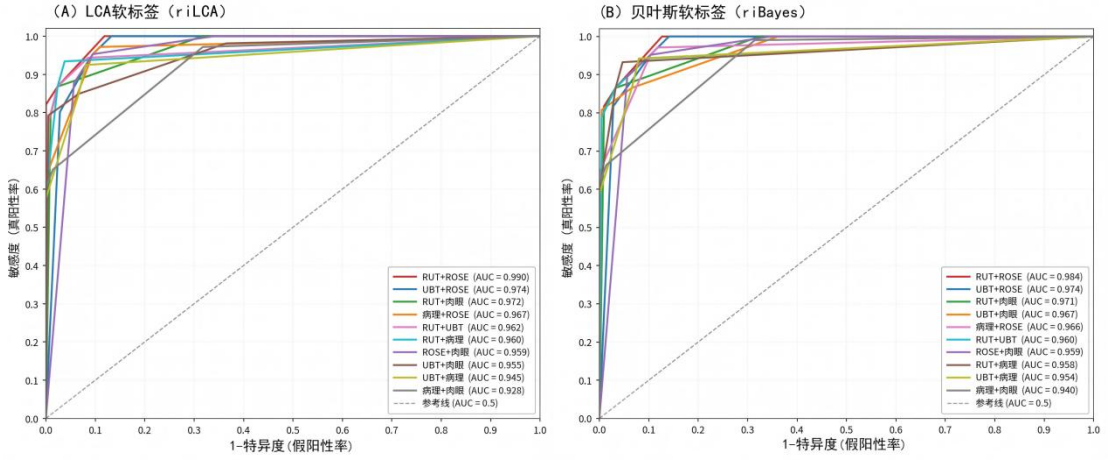
联合方案	串联 AND（两者均阳性才判阳）							并联 OR（任一阳性即判阳）							最优
	敏感度 Se		特异度 Sp		Youden 指数	PPV	NPV	敏感度 Se		特异度 Sp		Youden 指数	PPV	NPV	
RUT+ROSE	82.1%	(73.7－88.2)	**100.0%**	**(98.2－100.0)**	0.821	100.0%	91.7%	**100.0%**	**(96.5－100.0)**	88.2%	(83.1－91.8)	**0.882**	80.9%	100.0%	**并联**
UBT+ROSE	80.2%	(71.6－86.7)	**97.2%**	**(93.9－98.7)**	0.773	93.4%	90.7%	**100.0%**	**(96.5－100.0)**	86.7%	(81.5－90.7)	**0.867**	79.1%	100.0%	**并联**
RUT+肉眼	79.2%	(70.6－85.9)	**99.1%**	**(96.6－99.7)**	**0.783**	97.7%	90.5%	**100.0%**	**(96.5－100.0)**	68.2%	(61.7－74.2)	0.682	61.3%	100.0%	**串联**
病理+ROSE	63.2%	(53.7－71.8)	**100.0%**	**(98.2－100.0)**	0.632	100.0%	84.4%	**97.2%**	**(92.0－99.0)**	89.1%	(84.2－92.6)	**0.863**	81.7%	98.4%	**并联**
RUT+UBT	77.4%	(68.5－84.3)	**99.5%**	**(97.4－99.9)**	0.769	98.8%	89.7%	**94.3%**	**(88.2－97.4)**	91.5%	(86.9－94.5)	**0.858**	84.7%	97.0%	**并联**
RUT+病理	58.5%	(49.0－67.4)	**100.0%**	**(98.2－100.0)**	0.585	100.0%	82.7%	**93.4%**	**(87.0－96.8)**	96.2%	(92.7－98.1)	**0.896**	92.5%	96.7%	**并联**
ROSE+肉眼	87.7%	(80.1－92.7)	**94.3%**	**(90.3－96.7)**	**0.820**	88.6%	93.9%	**100.0%**	**(96.5－100.0)**	65.9%	(59.2－71.9)	0.659	59.6%	100.0%	**串联**
UBT+肉眼	79.2%	(70.6－85.9)	**99.5%**	**(97.4－99.9)**	**0.788**	98.8%	90.5%	**98.1%**	**(93.4－99.5)**	63.5%	(56.8－69.7)	0.616	57.5%	98.5%	**串联**
UBT+病理	57.5%	(48.0－66.5)	**100.0%**	**(98.2－100.0)**	0.575	100.0%	82.4%	**92.5%**	**(85.8－96.1)**	91.9%	(87.5－94.9)	**0.844**	85.2%	96.0%	**并联**
病理+肉眼	60.4%	(50.9－69.2)	**100.0%**	**(98.2－100.0)**	0.604	100.0%	83.4%	**97.2%**	**(92.0－99.0)**	68.2%	(61.7－74.2)	**0.654**	60.6%	98.0%	**并联**

表 15 贝叶斯软标签串联 (AND) vs 并联 (OR) 工作特性对照表

联合方案	串联 AND（两者均阳性才判阳）							并联 OR（任一阳性即判阳）							最优
	敏感度 Se		特异度 Sp		Youden 指数	PPV	NPV	敏感度 Se		特异度 Sp		Youden 指数	PPV	NPV	
RUT+ROSE	81.7%	(73.2－88.0)	**99.1%**	**(96.6－99.7)**	0.808	97.7%	91.7%	**100.0%**	**(96.4－100.0)**	87.3%	(82.2－91.1)	**0.873**	79.4%	100.0%	**并联**
UBT+ROSE	81.7%	(73.2－88.0)	**97.2%**	**(94.0－98.7)**	0.789	93.4%	91.6%	**100.0%**	**(96.4－100.0)**	85.9%	(80.6－90.0)	**0.859**	77.6%	100.0%	**并联**
RUT+肉眼	80.8%	(72.2－87.2)	**99.1%**	**(96.6－99.7)**	**0.798**	97.7%	91.3%	**100.0%**	**(96.4－100.0)**	67.6%	(61.1－73.5)	0.676	60.1%	100.0%	**串联**
UBT+肉眼	80.8%	(72.2－87.2)	**99.5%**	**(97.4－99.9)**	**0.803**	98.8%	91.4%	**100.0%**	**(96.4－100.0)**	63.8%	(57.2－70.0)	0.638	57.5%	100.0%	**串联**
病理+ROSE	64.4%	(54.9－73.0)	**100.0%**	**(98.2－100.0)**	0.644	100.0%	85.2%	**97.1%**	**(91.9－99.0)**	88.3%	(83.2－91.9)	**0.854**	80.2%	98.4%	**并联**
RUT+UBT	78.8%	(70.0－85.6)	**99.5%**	**(97.4－99.9)**	0.784	98.8%	90.6%	**94.2%**	**(88.0－97.3)**	90.6%	(85.9－93.8)	**0.848**	83.1%	97.0%	**并联**
ROSE+肉眼	89.4%	(82.0－94.0)	**94.4%**	**(90.4－96.7)**	**0.838**	88.6%	94.8%	**100.0%**	**(96.4－100.0)**	65.3%	(58.6－71.3)	0.653	58.4%	100.0%	**串联**
RUT+病理	59.6%	(50.0－68.5)	**100.0%**	**(98.2－100.0)**	0.596	100.0%	83.5%	**93.3%**	**(86.8－96.7)**	95.3%	(91.6－97.4)	**0.886**	90.7%	96.7%	**并联**
UBT+病理	58.7%	(49.0－67.6)	**100.0%**	**(98.2－100.0)**	0.587	100.0%	83.2%	**94.2%**	**(88.0－97.3)**	92.0%	(87.6－95.0)	**0.862**	85.2%	97.0%	**并联**
病理+肉眼	61.5%	(51.9－70.3)	**100.0%**	**(98.2－100.0)**	0.615	100.0%	84.2%	**99.0%**	**(94.8－99.8)**	68.5%	(62.0－74.4)	**0.676**	60.6%	99.3%	**并联**

注：加粗=该指标在两种规则中更优

图 17



注：每条曲线左下折点=串联(AND)工作点；右上折点=并联(OR)工作点。

表 16 LCA 软标签 DeLong 检验 — 两两 AUC 差异 p 值矩阵

	RUT+ROSE	UBT+ROSE	RUT+肉眼	病理+ROSE	RUT+UBT	RUT+病理	ROSE+肉眼	UBT+肉眼	UBT+病理	病理+肉眼

	RUT+ROSE	UBT+ROSE	RUT+肉眼	病理+ROSE	RUT+UBT	RUT+病理	ROSE+肉眼	UBT+肉眼	UBT+病理	病理+肉眼
RUT+ROSE	—	**0.023 ***	**2.60×10^{-4} ***	**0.019 ***	**0.010 ***	**0.007 ***	**8.91×10^{-4} ***	**0.003 ***	**0.002 ***	**7.22×10^{-6} ***
UBT+ROSE	**0.023 ***	—	0.886	0.457	0.379	0.366	0.127	0.112	**0.035 ***	**0.003 ***
RUT+肉眼	**2.60×10^{-4} ***	0.886	—	0.666	0.301	0.221	0.251	0.159	0.084	**0.002 ***
病理+ROSE	**0.019 ***	0.457	0.666	—	0.753	0.663	0.356	0.415	0.172	**0.007 ***
RUT+UBT	**0.010 ***	0.379	0.301	0.753	—	0.922	0.856	0.597	0.340	0.080
RUT+病理	**0.007 ***	0.366	0.221	0.663	0.922	—	0.937	0.756	0.415	0.052
ROSE+肉眼	**8.91×10^{-4} ***	0.127	0.251	0.356	0.856	0.937	—	0.771	0.407	**0.038 ***
UBT+肉眼	**0.003 ***	0.112	0.159	0.415	0.597	0.756	0.771	—	0.281	**0.035 ***
UBT+病理	**0.002 ***	**0.035 ***	0.084	0.172	0.340	0.415	0.407	0.281	—	0.256
病理+肉眼	**7.22×10^{-6} ***	**0.003 ***	**0.002 ***	**0.007 ***	0.080	0.052	**0.038 ***	**0.035 ***	0.256	—

表 17 贝叶斯软标签 DeLong 检验 — 两两 AUC 差异 p 值矩阵

	RUT+ROSE	UBT+ROSE	RUT+肉眼	UBT+肉眼	病理+ROSE	RUT+UBT	ROSE+肉眼	RUT+病理	UBT+病理	病理+肉眼
RUT+ROSE	—	0.169	**0.019 ***	0.054	0.071	**0.036 ***	**0.012 ***	**0.020 ***	**0.028 ***	**2.01×10^{-4} ***
UBT+ROSE	0.169	—	0.787	0.463	0.389	0.324	0.139	0.291	0.118	**0.011 ***
RUT+肉眼	**0.019 ***	0.787	—	0.691	0.684	0.294	0.325	0.183	0.259	**0.009 ***
UBT+肉眼	0.054	0.463	0.691	—	0.911	0.506	0.518	0.541	0.154	**0.031 ***
病理+ROSE	0.071	0.389	0.684	0.911	—	0.740	0.465	0.607	0.447	**0.033 ***
RUT+UBT	**0.036 ***	0.324	0.294	0.506	0.740	—	0.956	0.877	0.719	0.250
ROSE+肉眼	**0.012 ***	0.139	0.325	0.518	0.465	0.956	—	0.918	0.748	0.130
RUT+病理	**0.020 ***	0.291	0.183	0.541	0.607	0.877	0.918	—	0.846	0.220
UBT+病理	**0.028 ***	0.118	0.259	0.154	0.447	0.719	0.748	0.846	—	0.327
病理+肉眼	**2.01×10^{-4} ***	**0.011 ***	**0.009 ***	**0.031 ***	**0.033 ***	0.250	0.130	0.220	0.327	—

注：*标注 p < 0.05（联合评分整体效能差异显著）

四、讨论

本研究基于 317 例疑似 Hp 感染患者的同步五法对照数据，在无金标准框架下评估了 F-ROSE 与四种传统方法的诊断效能。以下从诊断性能、方法间对比、年龄段表现及方法学合理性四个方面展开讨论。

4.1 F-ROSE 诊断效能综合评价

在无金标准框架下，F-ROSE 展现出最优的综合诊断效能。其敏感性及阴性预测值（NPV）均为五种方法中最高，意味着 F-ROSE 极大地降低了漏诊风险，在以“排除诊断”为目的的场景中价值突出；同时其特异性保持在较高水平，K=2 与 K=3 两种 LCA 模型设定下排序不变，约登指数最佳。三种贝叶斯先验方案下，F-ROSE 的 Se、Sp 最大差异仅 0.97% 和 3.02%（九个文献先验参数中稳定性最

优）；三套互补 ROC 分析（以 LCA 软标签、贝叶斯软标签为参照的加权 ROC，以及以贝叶斯后验概率为评分的一致性 ROC）共同证实，F-ROSE 在两套软标签框架下加权 AUC 均为最高（riLCA 0.949、riBayes 0.945），且两套参照下 AUC 绝对差异均不超过 0.02；贝叶斯后验中位数 Se=0.948 与 EM-MLE 0.947 几乎完全重合（差异 0.1%）；提示 F-ROSE 的诊断效能高度稳健，不依赖于特定的模型算法或先验假设。

除了基础诊断效能，F-ROSE 更具备传统方法欠缺的两大临床优势：一是分级判读能力，本研究中多数阳性病例为低载量（"+"级）感染，提示 F-ROSE 对弱阳性样本极其敏感，弥补了定性方法易漏诊的不足；二是活/死菌鉴别能力，本研究发现活菌信号往往伴随显著的胃黏膜病变，这为评估感染活动度、指导根除治疗时机提供了独特的微生物学可视化证据。此外，其 25-30 分钟的出报告速度，真正实现了内镜检查的“现场快速评估（ROSE）”。

4.2 与传统方法的对比分析

与传统方法相比，F-ROSE 具有显著的方法学优势和互补性：

相较于病理检查：病理虽特异性极高，但敏感性不足（易受取材局灶性和主观判读影响），且耗时长**错误！不能识别的开关参数。**。F-ROSE 在同样使用活检标本的前提下，突破了病理学的时效瓶颈，并大幅提升了检出率。

相较于^{13}C-UBT：DeLong 加权检验证实 F-ROSE 显著优于^{13}C-UBT（riLCA P=0.0162，riBayes P=0.0212），F-ROSE 在诊断效能上全面占优。对于已有胃镜指征的患者，同次就诊完成 F-ROSE 检测，比额外安排易受药物/饮食干扰的^{13}C-UBT 更为高效、经济。

相较于 RUT：DeLong 检验显示 F-ROSE 与 RUT 的 AUC 差异在两套软标签下临界（riLCA P=0.0456 接近显著，riBayes P=0.2654 无显著差异），提示两者整体效能确实接近。但 F-ROSE 不依赖尿素酶代谢活性，能提供直观的形态学和分级证据，有效避免了低菌量或抑制状态下 RUT 的假阴性**错误！不能识别的开关参数。**。

相较于肉眼观察：肉眼判断的假阳性率极高，受主观经验影响大，本研究再次证实其不能单独作为 Hp 感染的确诊依据。

在两两联合策略中，RUT+F-ROSE 表现出最优的协同效应，RUT+F-ROSE 加权 AUC 在两套软标签下分别为 0.990 和 0.984，配对 DeLong 检验显示其在 riLCA 参照下显著高于其余 9 种两两联合方案（全部 P<0.05）。由于两者基于同次活检标本，无额外创伤，并联可实现极高的敏感性（适用于筛查），串联可实现极高的特异性（适用于确诊），是临床路径中最具可行性和高效性的双重检测方案。

4.3 各检测方法在不同年龄段和性别中的表现

本研究观察到≤40 岁组阳性率偏高的年轻化趋势，虽未达校正后显著性，但不同方法对年龄增长的敏感度差异值得关注。随年龄增长，五种方法阳性率均呈下降趋势，下降幅度分别为肉眼 20.2、病理 16.6、UBT 16.3、F-ROSE 15.1、RUT 14.2 个百分点，F-ROSE 降幅较小。值得注意的是，F-ROSE 在三个年龄组中均维持四种客观方法（除肉眼观察外）的最高阳性检出率，提示其在中老年人群中仍具备稳健的检出能力。这可能因为中老年人群常伴随萎缩、肠化等病变导致菌量降低，对依赖代谢活性的间接检测构成挑战**错误！不能识别的开关参数。**，而 F-ROSE 基于直接形态学识别，受年龄及胃黏膜退行性变的影响较小。此外，多重比较校正后，性别因素对五种方法的检出率均无统计学显著影响（包括校正前 P=0.017 的病理检查），提示性别不是 Hp 检测结果的主要混杂因素。

4.4 无金标准分析框架的方法学合理性

本研究框架亦为后续无金标准条件下的诊断试验评价提供了可复用的方法学模板。鉴于传统 Hp 检测方法均存在固有的假阳性或假阴性，强行指定任一“金标准”都会引入系统性偏倚（本研究 McNemar 检验已证实此点）。为此，本研究构建了“LCA—BLCM—DeLong 检验”三层互证的方法学体系。多重独立分析路径的结论高度收敛（LCA 与贝叶斯估计结果高度一致），不仅从方法学上坚实支撑了 F-ROSE 效能最优的核心结论，也为后续无金标准条件下的诊断试验评价提供了可复用的模板。

同时，在 LCA 类别数选择上，本研究并未机械追随单一信息准则（BIC 在 K=3 时最小），而是综合考量了研究估计目标的结构匹配性（Se/Sp 在数学定义上要求二分类潜变量）、临床决策的二分类属性、绝对拟合优度、分类清晰度及局部依赖性的可能干扰，最终选定 K=2。稳健性分析进一步显示，K=2 与 K=3 模

型对 86% 受试者(273/317)的分类完全一致,五种方法的 Se/Sp 排序在 K=3 的两种合并方案下与 K=2 高度一致，核心诊断结论不因类别数选择而改变。这一结构匹配优先于纯统计指标的决策框架，符合 Nylund 等 2007 **错误！不能识别的开关参数。**；Masyn 2013 **错误！不能识别的开关参数。**；Weller 等 2020 **错误！不能识别的开关参数。**的现代 LCA 模型选择建议，亦为后续无金标准诊断研究提供方法学参照。

4.5 研究的局限性与展望

本研究存在以下局限：(1)方法学设计的固有局限：本研究基于无金标准潜在类别模型推导真实感染状态,评价结果本质上反映的是方法间的共识性而非独立外部验证效能，这是本研究设计固有的方法学局限，提示结论外推时需审慎。(2)样本局限性:作为单中心研究且人群为具备胃镜指征的高潜在患病率(34.79%)群体，其结论向一般人群外推时需谨慎，未来需多中心大样本验证并积累独立的 Se/Sp 文献先验；(3)横断面设计：缺乏对患者根除治疗后的长期随访，F-ROSE 分级及活/死菌状态与预后、复发风险及耐药性的深层关联有待前瞻性队列研究证实；(4)方法独立性假设：本研究 K=3 模型的中间类剖面提示，F-ROSE 与胃镜白光直视观察之间可能存在局部依赖（内镜操作中医师对疑似病灶的口头沟通），这是 LCA 局部独立性假设的轻微违反。后续多中心研究应通过严格的盲法操作流程（如 F-ROSE 取样者与内镜判读者完全独立、不互通信息）进一步控制此偏倚；(5）卫生经济学：尚未评估 F-ROSE 初期设备投入与长期临床获益的成本-效益比。

基于上述局限性，后续研究方向包括：（1）开展多中心、大样本、前瞻性研究,进一步验证 F-ROSE 的诊断效能并积累其 Se/Sp 先验证据;(2)建立 F-ROSE 分级与根除治疗反应、长期复发风险的关联模型，挖掘其分级判读的临床决策价值;(3)开展卫生经济学评价,明确其在不同医疗资源配置下的成本-效益地位;（4）探索 F-ROSE 在 Hp 根除治疗后疗效评估、耐药性预测、合并其他消化道病变（如非典型增生、胃癌前病变）筛查中的扩展应用。

4.6 临床推荐应用路径

基于本研究证据，对临床 Hp 检测路径提出如下推荐：（1）对于已具备胃镜指征的患者，建议在胃镜活检同时开展 F-ROSE 检测作为首选快速诊断手段，可在同次就诊周期内（25–30 分钟）出具结果，显著缩短诊断—治疗启动间隔。（2）当临床需要更高诊断把握度（如根除治疗启动前的确诊、疑难病例复核）时，推荐采用 RUT+F-ROSE 的双重联合检测：以串联（AND）规则用于确诊（特异性 100%、PPV 100%），以并联（OR）规则用于排除诊断或筛查（敏感性 100%、NPV 100%），具体规则选择应根据临床决策目的（rule-in 或 rule-out）确定。（3）对于无胃镜指征、以筛查或随访为目的的人群，UBT 仍是合理的无创首选**错误！不能识别的开关参数。**，但需严格控制 PPI、抗生素停药时间。（4）胃镜白光直视判断由于特异性不足、主观性强，不应单独作为 Hp 感染的确诊依据，仅可作为辅助提示信息。（5）对于 F-ROSE 检测中观察到活菌信号（尤其伴"++"/"+++"分级且胃镜下显著病变）的患者，应警惕感染处于活动期，建议优先安排根除治疗并加强随访。

综上所述，F-ROSE 是一种快速、敏感、稳健且具有独特临床附加价值的 Hp 感染检测新方法，与 RUT 联合应用可形成兼具高敏感性与高特异性的最优诊断策略，有望成为 Hp 感染临床快速诊断的首选手段。本研究亦为无金标准条件下诊断试验效能评价提供了可复用的方法学框架。

[参考文献]

[1] Sugano K, Tack J, Kuipers E J, et al. Kyoto global consensus report on Helicobacter pylori gastritis[J]. Gut, 2015, 64(9): 1353-1367.

[2] Malfertheiner P, Megraud F, Rokkas T, et al. Management of Helicobacter pylori infection: the Maastricht VI/Florence consensus report[J]. Gut, 2022, 71(9): 1724-1762.

[3] Miftahussurur M, Yamaoka Y. Diagnostic Methods of Helicobacter pylori Infection for Epidemiological Studies: Critical Importance of Indirect Test Validation[J]. BioMed Research International, 2016, 2016: 4819423.

[4] Olaiya T F, Ajayi A, Smith S I. Current update on the diagnosis of Helicobacter pylori infection – 2024-2025[J]. Microbiota in Health and Disease, 2025, 7: e1438.

[5] 咸涛, 袁相恋, 张兴全, 等. 基于人工智能荧光成像识别幽门螺旋杆菌的方法及装置: 中国, CN202511164504.1[P]. 2025-12-02.

[6] 杜海莲, 田春燕, 赵利, 等. 人工智能与免疫荧光染色结合技术对支气管肺泡灌洗液病原学诊断的快速现场评价应用研究[J]. 现代检验医学杂志, 2026, 41(1): 170-174, 184.

[7] 宫颈液基细胞学人工智能辅助诊断数据集标注规范与质量控制专家共识(2022 版)编写组. 宫颈液基细胞学人工智能辅助诊断数据集标注规范与质量控制专家共识(2022 版)[J]. 中华病理学杂志, 2022, 51(12): 1205-1209.

[8] 陈洁, 李文生, 张巍. 人工智能辅助系统在宫颈液基细胞学分析中的应用价值研究[J]. 现代检验医学杂志, 2023, 38(5): 155-159.

[9] 张缨, 王晓露, 秦峰, 等. 免疫荧光染色检测胃黏膜活检标本中幽门螺杆菌和真菌感染的应用研究[J]. 胃肠病学, 2021, 26(1): 30-34.

[10] 陈丽雅, 王凤翔, 朱方超, 等. 免疫荧光染色在幽门螺杆菌检测中的应用价值[J]. 中国现代医生, 2022, 60(21): 42-45.

[11] Lauer B A, Reller L B, Mirrett S. Comparison of acridine orange and Gram stains for detection of microorganisms in cerebrospinal fluid and other clinical specimens[J]. Journal of Clinical Microbiology, 1981, 14(2): 201-205.

[12] Landis J R, Koch G G. The measurement of observer agreement for categorical data[J]. Biometrics, 1977, 33(1): 159-174.

[13] Mazeri S, Sargison N, Kelly R F, et al. Evaluation of the performance of five diagnostic tests for Fasciola hepatica infection in naturally infected cattle using a Bayesian no gold standard approach[J]. PLOS ONE, 2016, 11(8): e0161621.

[14] Hui S L, Walter S D. Estimating the error rates of diagnostic tests[J]. Biometrics, 1980, 36(1): 167-171.

[15] Albert P S, Dodd L E. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard[J]. Biometrics, 2004, 60(2): 427-435.

[16] Joseph L, Gyorkos T W, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard[J]. American Journal of Epidemiology, 1995, 141(3): 263-272.

[17] Nylund K L, Asparouhov T, Muthén B O. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study[J]. Structural Equation Modeling, 2007, 14(4): 535-569.

[18] Masyn K E. Latent class analysis and finite mixture modeling[M]//The Oxford Handbook of Quantitative Methods. Oxford University Press, 2013, 2: 551-611.

[19] Weller B E, Bowen N K, Faubert S J. Latent class analysis: A guide to best practice[J]. Journal of Black Psychology, 2020, 46(4): 287-311.

[20] Cheung A, Firestone S M. Bayesian latent class analysis when the reference test is imperfect[J]. Revue Scientifique et Technique (International Office of Epizootics), 2021, 40(1): 271-281.

[21] Ren S, Cai P, Liu Y, et al. Prevalence of Helicobacter pylori infection in China: A systematic review and meta-analysis[J]. J Gastroenterol Hepatol, 2022, 37(3): 464-470.

[22] Jambi L K. Systematic Review and Meta-Analysis on the Sensitivity and Specificity of 13C/14C-Urea Breath Tests in the Diagnosis of Helicobacter pylori Infection[J]. Diagnostics, 2022, 12(10): 2428.

[23] Omata F, Ohde S, Deshpande G, et al. Diagnostic Performance of Three Endoscopic Tests for Helicobacter pylori Infection: Systematic Review and Meta-analysis[J]. Am J Gastroenterol, 2015, 110(Suppl 1): S2452.

[24] Lee J G, Yoo I K, Yeniova A O, et al. The Diagnostic Performance of Linked Color Imaging Compared to White Light Imaging in Endoscopic Diagnosis of Helicobacter pylori Infection: A Systematic Review and Meta-Analysis[J]. Gut Liver, 2024, 18(3): 444-456.

[25] Gelman A, Rubin D B. Inference from Iterative Simulation Using Multiple Sequences[J]. Statistical Science, 1992, 7(4): 457-472.

[26] DeLong E R, DeLong D M, Clarke-Pearson D L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach[J]. Biometrics, 1988, 44(3): 837-845.

[27] 咸涛, 袁相恋, 何元林. 一种基于快速共聚焦的 Hp 药敏检测试剂及其在幽门螺杆菌检测中的应用: 中国, CN202610151102.6[P]. 2026-04-03.

[28] Wang X, Shu X, Li Q, et al. Prevalence and risk factors of Helicobacter pylori infection in Wuwei, a high-risk area for gastric cancer in northwest China: An all-ages population-based cross-sectional study[J]. Helicobacter, 2021, 26(4):e12810.

10.12201/bmr.202606.00055V1

[29] Dixon M F, Genta R M, Yardley J H, et al. Classification and grading of gastritis: the updated Sydney System. International Workshop on the Histopathology of Gastritis, Houston 1994[J]. The American Journal of Surgical Pathology, 1996, 20(10): 1161-1181.

[30] Wang Y K, Kuo F C, Liu C J, et al. Diagnosis of Helicobacter pylori infection: current options and developments[J]. World Journal of Gastroenterology, 2015, 21(40): 11221-11235.

[31] Lahner E, Vaira D, Figura N, et al. Role of noninvasive tests (13C-urea breath test and serology) in the diagnosis of Helicobacter pylori infection in patients with severe atrophic body gastritis[J]. The American Journal of Gastroenterology, 2004, 99(10): 1911-1917.

[32] Crowe S E. Helicobacter pylori Infection[J]. New England Journal of Medicine, 2019, 380(12): 1158-1165.

10.12201/bmr.202606.00055V1