Advances in Clinical Medicine 临床医学进展, 2026, 16(5), 3611-3616
Published Online May 2026 in Hans. https://www.hanspub.org/journal/acm
https://doi.org/10.12677/acm.2026.1652185

基于住院病历的结直肠息肉患病风险预测模型构建

曾　萍

珠海市中西医结合医院消化内科，广东　珠海

收稿日期：2026年4月28日；录用日期：2026年5月22日；发布日期：2026年5月29日

摘　要

目的：利用住院病历常规数据构建简洁、可解释的结直肠息肉患病风险预测模型，为临床提供无创、低成本的内镜筛查决策工具。方法：回顾性纳入2023年1月~2025年5月珠海市中西医结合医院80例首次接受结肠镜检查的住院患者，以病理确诊“是否存在息肉”为结局。通过ETL脚本自动提取入院24 h内人口学、症状、实验室及用药信息，采用LASSO (Least Absolute Shrinkage and Selection Operator)回归进行变量筛选以应对小样本量下的过拟合问题，并构建多因素Logistic回归模型。经共线性诊断与变量筛选后，以AUC、校准度及临床决策曲线评价模型性能。对缺失数据，采用多重插补法(m = 5)进行处理，并详细记录了缺失变量的分布与插补策略。结果：51例(63.75%)检出息肉。多因素分析最终保留4个变量：年龄、BMI、便血及癌胚抗原(CEA)。模型AUC = 0.92，Hosmer-Lemeshow P = 0.469，回归方程为Logit (P) = −25.42 + 0.11 × 年龄 + 0.73 × BMI + 2.39 × 便血 + 1.12 × CEA。结论：基于住院常规资料的4因子Logistic模型预测效能良好、校准度高，无需额外检测即可实现“一键式”风险计算，适合嵌入HIS系统辅助内镜排程，并可为结直肠癌一级预防提供可操作工具。

关键词

结直肠息肉，风险预测，住院病历，Logistic回归

Construction of Risk Prediction Model for Colorectal Polyps Morbidity Based on Hospitalized Medical Records

Ping Zeng

Department of Gastroenterology, Zhuhai Hospital of Integrated Traditional Chinese and Western Medicine, Zhuhai Guangdong

Received: April 28, 2026; accepted: May 22, 2026; published: May 29, 2026

文章引用：曾萍. 基于住院病历的结直肠息肉患病风险预测模型构建[J]. 临床医学进展, 2026, 16(5): 3611-3616.
DOI: 10.12677/acm.2026.1652185

Abstract

Objective: To construct a simple and interpretable risk prediction model for colorectal polyps by using routine data of inpatient medical records, and to provide a noninvasive and low-cost decision-making tool for endoscopic screening in clinic. Methods: From January 2023 to May 2025, 80 inpatients who received colonoscopy for the first time in Zhuhai Hospital of Integrated Traditional Chinese and Western Medicine were retrospectively included, and the pathological diagnosis was "whether there were polyps". The demographic, symptom, laboratory and medication information within 24 hours after admission was automatically extracted by ETL script. The variables were screened by LASSO (Least Absolute Shrinkage and Selection Operator) regression to deal with the over-fitting problem under small sample size, and a multi-factor Logistic regression model was constructed. After collinearity diagnosis and variable screening, the model performance was evaluated by AUC, calibration and clinical decision curve. The missing data are processed by multiple interpolation (m = 5), and the distribution and interpolation strategy of missing variables are recorded in detail. Results: Polyps were detected in 51 cases (63.75%). Multivariate analysis finally retained four variables: age, BMI, hematochezia and carcinoembryonic antigen (CEA). AUC = 0.092, Hosmer-Lemeshow P = 0.469, and the regression equation is logit (P) = −25.42 + 0.11 × age + 0.73 × BMI + 2.39 × hematochezia + 1.12 × CEA. Conclusion: The 4-factor Logistic model based on routine hospitalization data has good prediction efficiency and high calibration, and can realize "one-button" risk calculation without additional detection, which is suitable for being embedded in HIS system to assist endoscopic scheduling, and can provide an operational tool for primary prevention of colorectal cancer.

Keywords

Colorectal Polyps, Risk Prediction, Hospitalization Medical Records, Logistic Regression

1. 前言

结直肠息肉是结直肠癌重要的癌前病变，早期识别其患病风险对降低结直肠癌发病率和死亡率具有重要意义[1]。然而，目前我国结直肠息肉的筛查主要依赖于结肠镜检查，受限于人群依从性及医疗资源分布，早期发现率仍不理想[2]。随着医院信息化建设的推进，住院病历中积累了大量结构化与非结构化健康数据，涵盖患者既往史、实验室检查、影像学资料及用药记录等多维度信息，为构建风险预测模型提供了丰富的数据基础[3] [4]。基于临床住院数据构建结直肠息肉患病风险预测模型，不仅有助于提升早期识别效率，还可为临床决策提供客观依据[5]。本研究拟利用住院病历数据，构建一种高效、可解释的结直肠息肉风险预测模型，以期为精准筛查与个体化干预提供科学支持。

2. 资料与方法

2.1. 一般资料

本研究为回顾性单中心观察性研究。数据来源于医院住院电子病历系统。纳入 2023 年 1 月~2025 年 5 月期间因腹痛、便血、排便习惯改变或体检指征入院、首次接受完整结肠镜检查并留存完整病历的住院患者 80 例。纳排标准：① 年龄 18~85 岁；② 结肠镜抵达回盲部，检查描述及病理报告完整；③ 病历

中人口学、既往史、实验室、影像学、用药记录字段缺失率 < 10%。排除：① 既往结直肠息肉或肿瘤手术史；② 炎症性肠病、家族性腺瘤性息肉病、遗传性非息肉病性结直肠癌；③ 严重肝肾功能不全或妊娠期；④ 近 3 个月内服用 NSAIDs ≥ 7 d 或激素/免疫抑制剂。最终 80 例全部进入建模集，其中男性 39 例、女性 41 例，年龄 44~82 (58.01 ± 9.33)岁；BMI 16.82~33.73 (23.85 ± 3.43) kg·m^{-2}；合并高血压 26 例、2 型糖尿病 25 例、血脂异常 35 例；吸烟 20 例、饮酒 6 例。结肠镜下检出息肉 51 例(63.8%)，非息肉 29 例(36.25%)，息肉最大径 2~40.4 mm，病理提示管状腺瘤 45 例、绒毛成分 9 例、锯齿状病变 6 例。

2.2. 方法

通过医院电子病历系统导出 80 例患者首次入院 24 h 内生成的全部结构化及自由文本记录。使用 ETL 脚本自动抓取：① 人口学字段；② 主诉与现病史；③ 既往史、个人史、家族史；④ 首次实验室全血、生化、凝血、肿瘤标志物；⑤ 腹盆 CT/MRI 报告；⑥ 住院首次长期医嘱用药。自由文本采用正则匹配 + 人工复核方式补全，缺失值处理采用系统化流程：首先，对缺失率 > 10%的变量予以剔除；对于保留的变量，若存在缺失值，则采用多重插补法(Multiple Imputation, m = 5)进行填补，该方法通过链式方程(MICE)构建插补模型，能有效保留变量间的关联结构并减少插补偏差。异常值依据 3σ 原则与临床阈值双重截尾。以结肠镜病理“是否存在息肉”为二分类结局。

2.3. 观察指标

① 结局指标：结肠镜病理确诊的结直肠息肉(含管状腺瘤、绒毛腺瘤、锯齿状病变等任何组织学类型)。

② 候选预测因子：a. 人口学：性别、年龄、BMI；b. 生活方式：吸烟、饮酒；c. 既往史：高血压、糖尿病、血脂异常、慢性肝病、胆囊结石；d. 症状体征：腹痛、便血、排便频率、大便性状改变、体质量下降；e. 实验室：白细胞计数(WBC)、血红蛋白(Hb)、血小板计数(PLT)、空腹血糖(FPG)、糖化血红蛋白(HbA1c)、总胆固醇(TC)、癌胚抗原(CEA)；f. 用药：阿司匹林、氯吡格雷、他汀、PPI、口服降糖药。

2.4. 统计学方法

采用 SPSS25.0 与 R4.3.1 联合分析。计量资料以 $\bar{x} \pm s$ 表示，计数资料以例(%)表示，组间比较分别用独立样本 t 检验或 χ^2 检验；单因素分析后，将 P < 0.05 变量纳入共线性诊断(VIF > 5 剔除)，鉴于本研究样本量有限(n = 80)而初始候选变量较多，为避免传统逐步回归法在变量筛选过程中可能导致的过拟合与模型不稳定问题，我们选用 LASSO (Least Absolute Shrinkage and Selection Operator)回归进行变量筛选。LASSO 回归通过 L1 正则化压缩变量系数，能够自动进行变量选择并提升模型的泛化能力，特别适用于小样本、多变量的临床预测模型构建。继以 LASSO 回归初筛，最终行逐步 Logistic 回归建立预测模型。缺失值 < 10%的变量采用多重插补(m = 5)，>10%直接剔除；异常值依据 3σ 原则与临床界限双重截尾。显著性水平 α = 0.05 (双侧)。

3. 结果

3.1. 单因素分析

息肉组年龄、BMI、空腹血糖(FPG)及癌胚抗原(CEA)水平均显著高于非息肉组(P < 0.05)；男性、糖尿病、血脂异常既往史及便血症状亦在息肉组中占比更高(P < 0.05)。其余变量差异无统计学意义(P > 0.05)。如表 1 所示。

Table 1. Results of univariate analysis of study population
表 1. 研究人群单因素分析结果

候选预测因子	息肉组(n = 51)	非息肉组(n = 29)	统计量	P
a. 人口学				
性别(男/女)	32/19	9/20	χ^2 = 7.44	0.006
年龄(岁)	59.71 ± 9.16	55.03 ± 9.03	t =2.20	0.03
BMI ($kg \cdot m^{-2}$)	25.16 ± 3.12	21.54 ± 2.68	t = 5.48	<0.001
b. 生活方式				
吸烟[n (%)]	15 (29.41)	5 (17.24)	χ^2 =1.46	0.227
饮酒[n (%)]	6 (11.76)	0 (0.00)	χ^2 =3.69	0.055
c. 既往史				
高血压[n (%)]	18 (35.29)	8 (27.59)	χ^2 = 0.50	0.479
糖尿病[n (%)]	21 (41.17)	4 (13.80)	χ^2 = 6.453	0.01
血脂异常[n (%)]	26 (50.98)	8 (27.59)	χ^2 = 4.14	0.042
慢性肝病[n (%)]	12 (23.53)	8 (27.59)	χ^2 = 0.16	0.687
胆囊结石[n (%)]	10 (19.61)	3 (10.34)	χ^2 = 0.17	0.280
d. 症状体征				
腹痛[n (%)]	18 (35.29)	11 (37.93)	χ^2 = 0.06	0.814
便血[n (%)]	15 (29.40)	2 (6.90)	χ^2 = 5.60	0.02
排便频率改变[n (%)]	22 (43.14)	13 (44.83)	χ^2 = 0.02	0.884
大便性状改变[n (%)]	19 (37.25)	9 (31.03)	χ^2 =0.31	0.575
体质量下降[n (%)]	3 (5.88)	5 (17.24)	χ^2 = 2.65	0.104
e. 实验室				
WBC ($\times 10^9 \cdot L^{-1}$)	6.38 ± 1.84	5.64 ± 1.69	t = 1.81	0.075
Hb ($g \cdot L^{-1}$)	138.10 ± 14.85	133.31 ± 15.44	t = 1.35	0.182
PLT ($\times 10^9 \cdot L^{-1}$)	231.25 ± 60.52	248.83 ± 58.44	t = 1.28	0.207
FPG ($mmol \cdot L^{-1}$)	6.81 ± 2.32	5.48 ± 1.03	t =3.52	<0.001
HbA1c (%)	6.12 ± 1.40	5.81 ± 0.93	t = 1.19	0.239
TC ($mmol \cdot L^{-1}$)	5.21 ± 0.99	5.01 ± 0.82	t = 0.98	0.329
CEA ($\mu g \cdot L^{-1}$)	2.49 ± 1.19	1.57 ± 0.91	t = 3.85	<0.001
f. 用药				
阿司匹林[n (%)]	4 (7.84)	1 (3.45)	χ^2 = 0.61	0.435
氯吡格雷[n (%)]	0 (0.00)	1 (3.45)	χ^2 =1.78	0.182
他汀[n (%)]	6 (11.76)	2 (6.90)	χ^2 = 0.49	0.485
PPI [n (%)]	43 (84.31)	18 (62.07)	χ^2 = 5.05	0.025
口服降糖药[n (%)]	11 (21.57)	3 (10.34)	χ^2 = 1.61	204

注：连续变量以均数 ± 标准差表示，分类变量以例数(%)表示；统计量为 t 或 χ^2 值。

3.2. 多因素 Logistic 回归结果

通过 LASSO 回归(10 折交叉验证选择最优 λ 值)从单因素分析有意义的变量中筛选预测因子，最终年龄、BMI、便血和 CEA 被选中并纳入多因素 Logistic 回归模型。经共线性诊断及上述筛选，年龄、BMI、便血和 CEA 进入方程。结果显示，年龄每增加 1 岁，患病风险增加 11% (OR = 1.11, 95% CI: 1.02~1.22)；便血使风险升高约 10.9 倍(OR =10.92, 95% CI: 1.72~69.27)；CEA 每增加 1 μg·L^{-1}，风险增加 223% (OR = 3.23, 95% CI: 1.46~7.14)。模型整体 $\chi^2 = 54.16$，$P < 0.001$，AUC 达 0.92，Hosmer-Lemeshow 检验 $P = 0.64$，预测概率与实际观测吻合良好。构建的结直肠息肉风险预测模型如下：Logit (P) = −25.42 + 0.11 × 年龄 + 0.73 × BMI + 2.39 × 便血 + 1.12 × CEA；将上述 4 项指标取值代入方程即可获得个体患息肉的预测概率 $P = 1/(1 + e^{(-Logit(P))})$。如表 2 所示。

Table 2. Multivariate Logistic regression results of colorectal polyps risk (n = 80)
表 2. 结直肠息肉患病风险多因素 Logistic 回归结果(n = 80)

变量	β	SE	Wald χ^2	P	OR	95% CI	VIF
年龄	0.11	0.05	5.58	0.018	1.11	1.02~1.22	1.20
BMI	0.73	0.19	14.12	<0.001	110	1.41~3.01	1.02
便血	2.39	0.94	6.44	0.011	10.92	1.72~69.27	1.11
CEA	1.17	0.40	8.41	0.004	3.23	1.46~7.14	1.11
常数项	−25.42	6.58	14.91	<0.001	0.00	—	—

注：模型 $\chi^2 = 54.16$，df = 4，$P < 0.001$；AUC = 0.922；Hosmer-Lemeshow $\chi^2 = 6.107$，df = 8，$P = 0.635$。

4. 讨论

4.1. 模型预测效能与临床价值

本研究基于常规住院病历信息构建的 4 因素 Logistic 模型 AUC 达 0.92，高于既往仅依赖年龄、性别及家族史的传统评分(AUC 0.65~0.72) [6]，提示将实验室与症状变量纳入可显著提升区分度。模型 Hosmer-Lemeshow P = 0.635，校准度良好，适合在住院患者中开展电子化的“一键式”风险计算，为内镜资源分配提供量化依据。

4.2. 关键变量的生物学解释

年龄、BMI、糖尿病、空腹血糖与 CEA 均与胰岛素抵抗及低度慢性炎症相关，可协同促进息肉发生[7]。便血作为首发症状，本研究其 OR 达 10.92，与刘波等[8]在 1351 例儿童息肉研究中报告的“以便血为主要表现且息肉直径 ≥2 cm 时出血风险显著增高”结果相符，提示对住院患者任何程度的便血都应保持高度警觉。

4.3. 建模策略的可靠性

研究采用 ETL + 正则匹配双轨提取，字段缺失率 <10%，并通过多重插补降低信息偏倚；变量筛选采用 LASSO 回归，该方法通过压缩系数有效避免了小样本量下的过拟合问题，提升了模型的稳健性与泛化能力[9]；共线性诊断 VIF 均 <1.25，避免系数膨胀。逐步法保留 4 个变量，既保证 parsimony，又使模型易于在 HIS 系统落地，符合“轻量级”临床预测工具开发理念[9]。

4.4. 与国内外研究异同

程军等[10]对磺脲类降糖药与抗菌药物的潜在不良相互作用进行处方挖掘，证实利用常规电子病历

数据即可实现风险信号识别，其“二次数据利用”理念与本研究一致。不同于 Deiss-Yehiely 等[11]将多靶点粪便 DNA (mt-sDNA)或 CT 结肠成像(CTC)等分子/影像组学指标纳入模型(虽 AUC 可达 0.90 左右，但需额外设备与基因检测)，本模型变量均来自入院 24 h 内常规检验与问诊信息，无需新增成本，更适合在住院 HIS 系统中快速落地与推广。

4.5. 局限性

首先，单中心回顾性设计可能存在选择偏倚；其次，未纳入腹盆 CT 影像量化指标及粪便潜血结果，可能低估部分高危患者；再次，样本量仅 80 例，虽满足“事件数/变量 ＞10”经验法则[12]，但仍需外部大样本验证。后续将开展多中心前瞻性研究，并探索机器学习算法对非线性关系的捕获能力。

4.6. 结论与展望

本研究证实，基于住院常规资料的年龄、BMI、便血及 CEA 四因素模型具有较好的预测与校准能力，可在内镜排程软件中嵌入，实现高风险病例的自动提醒。未来将进一步优化变量阈值，并开发网页版计算器，为结直肠癌一级预防提供可操作的工具。

声　明

本研究获得珠海市中西医结合医院伦理委员会批准(2025-03-063-E02)。

参考文献

[1] 农云翠, 黄小知, 黄灵旭, 等. 结直肠息肉发生的影响因素分析[J]. 广西医学, 2025, 47(7): 962-967.

[2] 王人杰, 张晓兰, 蔡继东, 等. 结直肠息肉的规范化诊疗[J]. 中华胃肠外科杂志, 2024, 27(6): 583-590.

[3] 中华医学会消化病学分会医工交叉协作组. 结直肠息肉门诊管理专家共识(2025, 成都) [J]. 中华消化内镜杂志, 2025, 42(5): 337-347.

[4] 张庆林, 郑雯, 殷刚刚, 等. 胆囊息肉对结直肠息肉提示价值的相关性研究[J]. 中华消化内镜杂志, 2025, 42(3): 223-228.

[5] 胡堃苗, 毕玉珍, 余振华, 等. 结直肠息肉内镜治疗后迟发性出血的相关危险因素及风险预测模型构建[J]. 浙江创伤外科, 2025, 30(2): 315-318.

[6] 吕莹莹, 朱炳喜. 结直肠息肉高危人群早期筛查评分模型的建立[J]. 医学研究杂志, 2019, 48(8): 132-136.

[7] 梁晗. 2 型糖尿病与结直肠息肉相关性的临床研究[D]: [硕士学位论文]. 开封: 河南大学, 2021.

[8] 刘波, 张慧华, 张慧晖, 等. 儿童结直肠息肉 1351 例的临床特征及内镜下治疗效果分析[J]. 中国当代儿科杂志, 2022, 24(4): 354-359.

[9] 闫明海, 赵延延, 刘鑫, 等. 基于回归或机器学习方法的个体预后或诊断的多变量预测模型透明报告(TRIPOD + AI)解读[J]. 中华内科杂志, 2025, 64(1): 4-10.

[10] 程军, 汪龙, 张冠军, 等. 磺脲类降糖药与抗菌药物潜在不良药物相互作用的处方分析[J]. 医药导报, 2022, 41(5): 708-712.

[11] Deiss-Yehiely, N., Graffy, P.M., Weigman, B., Hassan, C., Matkowskyj, K.A., Pickhardt, P.J., *et al.* (2022) Detection of High-Risk Sessile Serrated Lesions: Multitarget Stool DNA versus CT Colonography. *American Journal of Roentgenology*, **218**, 670-676. https://doi.org/10.2214/ajr.21.26719

[12] 孙亚清. Logistic 回归样本量确定所需自变量事件数的模拟研究[D]: [硕士学位论文]. 广州: 南方医科大学, 2016.