

基于临床数据的胃癌筛查模型研究

杨荣¹, 陈誉², 高红梅¹, 陈先来^{3,4}

1. 中南大学湘雅医院, 湖南 长沙 410078; 2. 中南大学湘雅医学院, 湖南 长沙 410013; 3. 中南大学信息安全与大数据研究院, 湖南 长沙 410083; 4. 中南大学医疗大数据应用技术国家工程实验室, 湖南 长沙 410083

【摘要】目的:利用临床数据,通过机器学习建立辅助筛选模型,以提高胃癌早期诊断水平。**方法:**以5 585例胃癌(ICD编码为C16*,A组)患者为研究对象,并从57 657例非胃部恶性肿瘤(ICD编码为C*,除C16*外)中随机选择6 000例(B组),从47 225例健康体检者中随机选择6 000例非恶性肿瘤(C组),作为对照。从临床数据中抽取人口学(性别、年龄)、实验室检测(血常规检测、血脂/肝功能、肿瘤相关标志物、Hp等)等信息。利用Pearson相关性分析,对各指标与诊断之间的相关性进行分析。采用独立样本 t 检验,检测各指标的组间差异性。选择性别、年龄、癌胚抗原(CEA)、粪隐血(FOB)等53项指标作为决策变量,采用决策树算法C5.0,建立胃癌辅助筛查模型。**结果:**年龄、CEA、CA153等指标与胃癌显著相关($P<0.05$)。在A组-B组、B组-C组、A组-C组中,存在组间差异性的指标不相同。通过数据挖掘,得到了包含51条规则的胃癌筛查模型。模型中重要性位于前10的指标依次为CA199、CA153、CEA等。对于训练集、测试集,模型的准确率分别为89.58%、89.14%,曲线下面积为0.809。**结论:**通过临床数据分析,可以确定胃癌早期诊断的重要指标。利用数据挖掘方法,基于临床数据可以建立胃癌筛查辅助模型,对于胃癌筛查具有良好的辅助价值。

【关键词】胃肿瘤;疾病筛查模型;临床数据;决策树

【中图分类号】R319

【文献标志码】A

【文章编号】1005-202X(2019)09-1095-08

Clinical data-based model for gastric cancer screening

YANG Rong¹, CHEN Yu², GAO Hongmei¹, CHEN Xianlai^{3,4}

1. Xiangya Hospital, Central South University, Changsha 410078, China; 2. Xiangya School of Medicine, Central South University, Changsha 410013, China; 3. Information Security and Big Data Research Institute, Central South University, Changsha 410083, China; 4. National Engineering Laboratory for Medical Big Data Application Technology, Central South University, Changsha 410083, China

Abstract: Objective To establish an auxiliary screening model based on clinical data and machine learning for improving the early diagnosis of gastric cancer. **Methods** A total of 5 585 cases of gastric cancer (ICD code: C16*, group A) were selected as research subjects. In addition, 6 000 cases (group B) from 57 657 cases of non-gastric malignant tumors (ICD code: C*, except C16*) and 6 000 cases of non-malignant tumors (group C) from 47 225 healthy persons were randomly selected as controls. Demographical information (gender, age), laboratory tests (routine blood test, blood lipid/liver function, tumor-related markers, Hp, etc.) were extracted from clinical data. Pearson's correlation analysis was used to analyze the relationship between each indicator and diagnosis; and independent sample t test was performed for detecting the differences in indicators among different groups. A total of 53 indicators such as gender, age, carcinoembryonic antigen (CEA), fecal occult blood were selected as decision variables. An auxiliary model was established for gastric cancer screening by decision tree algorithm C5.0. **Results** The indicators such as age, CEA and CA153 were significantly correlated with gastric cancer ($P<0.05$). For the inter-group of group A and B, group B and C, group A and C, the indicators with inter-group differences were different. A model with 51 rules for gastric cancer screening was obtained by data mining. The top 10 indicators ranked by importance in the model were as follow: CA199, CA153, CEA, etc. The accuracy of the model was 89.58% for training set and 89.14% for test set. The area under curve was 0.809 for the model. **Conclusion** Through the analysis of clinical data, the important indicators for the early diagnosis of gastric cancer can be determined. An auxiliary model for gastric cancer screening can be established based on clinical data using data mining. The established model has excellent assistant value for gastric cancer screening.

Keywords: stomach neoplasms; model for disease screening; clinical data; decision tree

【收稿日期】2019-06-02

【基金项目】国家重点研发计划“精准医学研究”重点专项(2016YFC0901705);国家自然科学基金(13BTQ052)

【作者简介】杨荣,主管护师,主要研究方向:胃肠信息学,E-mail: cxlyr0576@163.com

【通信作者】高红梅,副主任护师,主要研究方向:临床护理学,E-mail: gaohongmei50@163.com

前言

降低胃癌的发病率和死亡率是亟待解决的重大公共卫生问题。2018年前9个月,美国胃癌患者死亡11 433例^[1]。2015年我国胃癌发病例数约为67.9万,其发病例数仅次于肺癌(73.3万例)^[2]。胃癌的发生发展是一个多因素参与、多步骤演变的复杂病理过程,包括人口学、感染、生活饮食、遗传和环境等^[3]。年龄、性别是胃癌的重要人口学因素,随着年龄增长,胃癌发病率和死亡率亦越来越高,男性的胃癌发病率和死亡率均高于女性^[4]。高盐饮食^[5]、腌熏煎油炸食品^[3]与胃癌关系密切。IARC、WHO均将Hp归为I类致癌因子,Hp感染与胃癌的发生存在相关性^[6-7]。流行病学资料显示部分胃癌有家族聚集倾向^[8]。精神心理社会因素(如精神刺激或抑郁)、免疫因素等可能与胃癌发生有一定关联^[9]。其它致病风险包括糖尿病^[10]、慢性萎缩性胃炎^[11]等。

胃癌的生存率与肿瘤进展情况密切相关,早期胃癌5年存活率可达90%^[12],而进展期胃癌5年存活率低于30%^[13]。然而,绝大多数胃癌早期症状不明显,与胃炎、胃溃疡等慢性胃部疾病相混淆,很难早期发现,容易导致漏诊和误诊。目前,胃癌的临床诊断方法大致有以下几类:(1)实验室检查。包括:①血常规检查。血红蛋白、红细胞、红细胞压积等指标下降,在早期胃癌筛查中有一定的价值。②肿瘤标志物测定、免疫学检查。如血清胃蛋白酶原I及胃蛋白酶原I/II之比^[14],CEA、CA199、CA125、CA724等^[15]癌类相关抗原及单克隆抗体检测,对胃癌的诊断具有一定价值,但这些检查的假阳性与假阴性均较高,特异性不强。③大便潜血试验。持续性大便潜血阳性,对胃癌的诊断有参考价值,可以为发现胃癌提供线索^[16-17]。(2)X线。如果X线钡餐检查发现可疑病变,则进一步进行内镜检查。(3)纤维内窥镜检查。它是胃癌诊断最直接准确有效的诊断方法,但依赖于设备和内镜医师资源,普通小型医疗机构难以开展,且费用相对较高,具有一定痛苦,患者接受程度较差,难以用于大规模胃癌筛查。(4)B超。通过超声检查,可了解胃部周围实质性脏器转移,但对于早期胃癌的诊断价值不大。(5)CT检查。通过CT检查可以了解胃肿瘤侵犯情况,以及与周围脏器关系,有无切除可能^[18]。此外,呼气的化学物质分析也对诊断胃癌有价值^[19]。以上各类诊断方法,存在特异性不强等问题,需要联合考虑。此外,胃癌的筛查还需要考虑危险因素。通过数据挖掘,可以从大量信息中提取有用的医学规则,建立疾病诊断辅助模型,

为疾病的筛查、诊断、预后分析提供支撑。本研究拟以临床数据为基础,尝试利用决策树算法,建立胃癌筛查模型,从中提取胃癌的筛查规则。目的在于通过构建良好的筛查模型,辅助胃癌早期诊断,提高胃癌的早期诊断水平。

1 对象与方法

1.1 数据收集

本研究选择胃癌患者作为研究对象,并以非胃部恶性肿瘤及健康体检者作为对照。研究对象的选择遵循以下原则:(1)胃癌患者第一出院诊断ICD编码为C16及其下位类;(2)非胃部恶性肿瘤患者第一出院诊断ICD编码为C及其下位类(C16及其下位类除外);(3)非恶性肿瘤者为无肿瘤诊断史体检者。数据来源于长沙某3所三甲医院,其中,肿瘤患者数据来源于临床医疗数据,非肿瘤患者数据来源于健康体检数据。基于以上原则,从住院病人中随机抽取胃癌患者5 585例[男3 597例,女1 966例,性别不详22例,年龄(56.96±11.81)岁],作为胃癌组(A组);从57 657例非胃部恶性肿瘤住院患者中随机抽取6 000例[男3 402例,女2 565例,性别不详33例,年龄(53.63±14.88)岁],作为非胃部恶性肿瘤组(B组);从42 275例健康体检者中随机抽取6 000例[男5 170例,女830例,年龄(40.57±12.51)岁],作为非恶性肿瘤组(C组)。

本研究数据收集范围的确定遵循以下原则:(1)已有研究显示可能的胃癌相关因素;(2)数据可以获取;(3)相关检测为常规经济性检测,而非昂贵的特殊检测;(4)相关检测为非侵入性或低侵入性。根据这些原则,主要收集人口学信息(性别、年龄)和实验室检测指标(血常规检测、血脂/肝功能、肿瘤相关标志物、Hp等),见表1。

1.2 数据预处理

数据来源于多家医疗机构,异构性强,存在名称不一致、值阈不一致、数据冗余、空缺严重等质量问题。对此,采取用LOINC统一指标、用归一化方法统一值阈、用相似性度量消除冗余、删除缺失率大于90%的指标等措施,对数据进行处理。预处理后数据如表2所示。

1.3 特征指标的选取

为提高筛查模型的鲁棒性,使用Pearson双变量相关性分析(双尾),对各指标与诊断(Diagnosis,非恶性肿瘤取0,非胃部恶性肿瘤取1,胃癌取2)之间相关性进行分析,识别出重要特征变量。与诊断显著相关的指标如表3所示。

表 1 所采集的实验室检测指标
Tab.1 Collected laboratory test indicators

检测指标	缩写代码	检测指标	缩写代码
ABO 血型系统	ABO	脂蛋白 α 测定	LPa
甲胎蛋白	AFP	淋巴细胞绝对值	Lym_Cnt
血清白蛋白	ALB	淋巴细胞百分比	Lym_Per
白球蛋白比值	ALB_2_GLB	平均红细胞血红蛋白含量	MCH
ApoA1/ApoB	ApoA_2_ApoB	平均红细胞血红蛋白浓度	MCHC
载脂蛋白 A1(apoA1)	ApoA1	平均红细胞体积	MCV
载脂蛋白 B(apoB)	ApoB	单核细胞绝对值	Mono_Cnt
活化部分凝血活酶时间	APTT_sec	单核细胞百分比	Mono_Per
嗜碱细胞绝对值	Baso_Cnt	血小板平均体积	MPV
嗜碱细胞百分比	Baso_Per	中性粒细胞绝对值	Neu_Cnt
血清补体 C3	C3	中性粒细胞百分比	Neu_Per
血清补体 C4	C4	大血小板比率	P_LCR
糖类抗原 125	CA125	血小板比积	PCT
糖类抗原 15-3	CA153	血小板分布宽度	PDW
糖类抗原 19-9	CA199	胃蛋白酶原 I	PG1
糖链抗原 242	CA242	胃蛋白酶原比值	PG1_2_PG2
糖类抗原 72-4	CA724	胃蛋白酶原 II	PG2
癌胚抗原	CEA	血小板计数	PLT
C-反应蛋白(CRP)	CRP	凝血酶原时间	PT
D-二聚体	D_Dimer	红细胞计数	RBC
直接胆红素	DBIL	红细胞分布宽度(CV)	RDW_CV
嗜酸细胞绝对值	Eos_Cnt	红细胞分布宽度(SD)	RDW_SD
嗜酸细胞百分比	Eos_Per	网织红细胞绝对值	RET_Cnt
大便潜血试验	FOB	网织红细胞百分比	RET_Per
血清球蛋白	GLB	Rh 血型鉴定	RH
血红蛋白	Hb	总胆汁酸	TBA
红细胞压积	HCT	总胆红素	TBIL
血清同型半胱氨酸	HCY	总胆固醇	TC
高密度脂蛋白胆固醇	HDL_C	甘油三酯	TG
高密度胆固醇比总胆固醇	HDL_C_2_TC	血清总蛋白	TP
幽门螺杆菌抗体	Hp	凝血酶时间	TT
超敏 C 反应蛋白	hs_CRP	25-羟基维生素 D	VitD_25_OH
乳酸脱氢酶	LDH_L	白细胞计数	WBC
低密度脂蛋白胆固醇	LDL_C	-	-

从表 3 可见,与诊断具有显著相关性($P<0.05$)的指标包括人口学因素(性别、年龄)、血液生化指标(如 ALB、TC、TG)、血常规检查指标(如 Hb、HCT、Lym_Per)、幽门螺杆菌感染(*Hp*)、胃蛋白酶原(PG1、PG2、PG1_2_PG2)以及癌类相关抗原(CA153、CA199)等。为了解各指标的组间差异显著性,进一

表2 预处理后的数据示例

Tab.2 Examples of data after preprocessing

序号	性别	年龄/岁	AFP	ALB
1	M	71	4.59	35.7
2	M	70	0.66	36.8
3	M	48	1.33	35.0
4	M	73	NULL	23.5
5	F	47	2.08	26.0
6	M	68	12.18	23.0
.....

步支持特征变量选取,对各指标进行组间独立样本 t 检验。结果发现,存在组间差异显著性的指标如表4所示。

1.4 胃癌筛查决策树模型的构建

建模变量的选取主要考虑以下因素:①与胃癌诊断的相关性;②组间差异显著性;③现有文献报道或临床实践情况。据此选取以下53项指标用于建模:性别、年龄、RH、AFP、ALB、ALB_2_GLB、APTT、Baso_Cnt、Baso_Per、CA125、CA153、CA199、CEA、DBIL、D_Dimer、Eos_Cnt、Eos_Per、FOB、GLB、Hb、HCT、HDL_C、HDL_C_2_TC、Hp、LDH_L、LDL_C、

表3 与诊断具有显著相关的指标

Tab.3 Indicators significantly correlated with diagnosis

指标	Pearson系数	P值	指标	Pearson系数	P值
性别	.161	0	MCV	-.057	0
年龄	.448	0	Mono_Cnt	.019	0.010
ABO	-.193	0	Mono_Per	.033	0
RH	.027	0.018	MPV	-.136	0
AFP	-.050	0.001	Neu_Cnt	.215	0
ALB	-.672	0	Neu_Per	.343	0
ALB_2_GLB	-.457	0	P_LCR	-.163	0
APTT	.045	0	PCT	.146	0
Baso_Per	-.069	0	PDW	-.200	0
CA153	-.057	0.002	PG1	.285	0
CA199	-.039	0.024	PG1_2_PG2	-.145	0
Eos_Cnt	.020	0.015	PG2	.477	0
FOB	.213	0	PLT	.189	0
Hb	-.616	0	PT	.065	0
HCT	-.573	0	RDW_CV	.263	0
HDL_C	-.366	0	RDW_SD	.035	0.039
HDL_C_2_TC	-.175	0	TBA	.033	0
Hp	.421	0	TBIL	-.063	0
LDL_C	-.073	0	TC	-.269	0
Lym_Cnt	-.141	0	TG	-.144	0
Lym_Per	-.454	0	TP	-.560	0
MCH	-.218	0	WBC	.040	0
MCHC	-.324	0	-	-	-

Lym_Cnt、Lym_Per、MCH、MCHC、MCV、Mono_Cnt、Mono_Per、MPV、Neu_Cnt、Neu_Per、P_LCR、PCT、PDW、PG1、PG1_2_PG2、PG2、PLT、PT、RBC、

RDW_CV、TBA、TBIL、TC、TG、TP、TT、WBC。

疾病筛查属于典型的分类问题,适合采用决策树算法建模。因此,先要对数值型数据进行离散处理。

表4 存在组间差异显著性的指标
Tab.4 Indicators with significant inter-group differences

组别	指标
A组 vs B组	性别、年龄、ABO、RH、AFP、ALB、ALB_2_GLB、CA125、CA153、CA199、DBIL、Eos_Cnt、Eos_Per、 FOB、GLB、Hb、HCT、HDL_C、HDL_C_2_TC、Lym_Cnt、Lym_Per、MCH、MCHC、MCV、MPV、 Neu_Cnt、Neu_Per、P_LCR、PCT、PG1_2_PG2、PLT、PT、RDW_CV、RDW_SD、TBA、TBIL、TC、 TG、TP、TT、WBC
A组 vs C组	性别、年龄、AFP、ALB、ALB_2_GLB、APTT_sec、Baso_Cnt、Baso_Per、CA125、CA199、CEA、 D_Dimer、DBIL、FOB、GLB、Hb、HCT、HDL_C、HDL_C_2_TC、Hp、LDH_L、LDL_C、Lym_Cnt、 Lym_Per、MCH、MCHC、MCV、Mono_Cnt、Mono_Per、MPV、Neu_Cnt、Neu_Per、P_LCR、PCT、 PDW、PG1、PG1_2_PG2、PG2、PLT、PT、RDW_CV、TBA、TBIL、TC、TG、TP、TT、WBC
B组 vs C组	性别、年龄、AFP、ALB、ALB_2_GLB、APTT_sec、Baso_Per、CEA、D_Dimer、DBIL、Eos_Cnt、 Eos_Per、FOB、GLB、Hb、HCT、HDL_C、HDL_C_2_TC、Hp、LDH_L、LDL_C、Lym_Cnt、Lym_Per、 MCH、MCHC、MCV、Mono_Cnt、Mono_Per、MPV、Neu_Cnt、Neu_Per、P_LCR、PDW、PG1、PG1_2_PG2、PG2、PT、RBC、RDW_CV、TBA、TC、TG、TP、TT、WBC

对于既有上限又有下限的变量,离散为正常(N)、偏低(L)、偏高(H)3种取值,如球蛋白(GLB);对于只有上限的变量,离散为正常(N)、偏高(H),如癌胚抗原(CEA)。参照WHO方法,年龄离散为青年及以下(<45岁)、中年(45~59岁)、老年及以上(≥60岁)。

每组按照6:4随机划分为训练集和测试集。本研究的目的在于确定是否有患胃癌(Diagnosis=2)风险,故将C组(Diagnosis=0)、B组(Diagnosis=1)合并为一组,作为非胃癌组(合并后Diagnosis=0)。经过划分、合并处理后,训练集包含C组记录3 580例、B组记录3 584例、A组记录3 340例;测试集包含C组记录2 420例、B组记录2 416例、A组记录2 245例。

本研究涉及的数据量大,考虑的变量(指标)多,故选择C5.0作为筛查模型构建算法。在模型构建时,设置目标变量为诊断(Diagnosis),前面所确定的指标作为决策变量。修剪严重性设置为75,每个子分支的最小记录数设为20,使用分区数据,并使用全局修剪方式,构建决策树模型。

2 结果

2.1 胃癌筛查模型与决策规则

从所得模型(图1)可以发现,从中可提取51条决策规则用于胃癌筛查。糖类抗原CA153是识别胃癌的首选指标,其次是糖类抗原CA199、癌胚抗原CEA,血常规指标HCT、RDW_CV、Baso_Cnt、Lym_Per以及胃蛋白酶原对于识别胃癌也有重要价值。

2.2 指标的重要性

在筛查模型中,重要性排前10的指标依次为CA199、CA153、CEA、HCT、RDW_CV、Baso_Cnt、Lym_Per、PG1_2_PG2、PG2、DBIL,如表5所示。

```
CA153 = [ Pattern: 0 ] (8,773)
CA199 = [ Pattern: 0 ] (8,522)
CEA = [ Pattern: 0 ] (7,732)
HCT in [ "L" ] [ Pattern: 0 ] (2,747)
.....
HCT in [ "N""H" ] [ Pattern: 0 ] => 0.0 (4,985; 0.957)
CEA = Z [ Pattern: 2 ] (674)
RDW_CV in [ ""L" ] [ Pattern: 2 ] => 2.0 (385; 0.919)
RDW_CV in [ "N""H" ] [ Pattern: 0 ] => 0.0 (289; 0.962)
CEA = H [ Pattern: 2 ] => 2.0 (116; 0.862)
CA199 = Z [ Pattern: 2 ] (205)
PG1_2_PG2 in [ ""L" ] [ Pattern: 2 ] => 2.0 (157; 0.987)
PG1_2_PG2 in [ "N""H" ] [ Pattern: 0 ] => 0.0 (48; 0.771)
CA199 = H [ Pattern: 2 ] => 2.0 (46; 0.935)
CA153 = Z [ Pattern: 2 ] (1,718)
PG1_2_PG2 in [ ""L" ] [ Pattern: 2 ] => 2.0 (1,601; 0.999)
PG1_2_PG2 in [ "N" ] [ Pattern: 2 ] (117)
Lym_Cnt in [ "" ] [ Pattern: 2 ] => 2.0 (0)
Lym_Cnt in [ "L""N" ] [ Pattern: 2 ] => 2.0 (103; 0.631)
Lym_Cnt in [ "H" ] [ Pattern: 0 ] => 0.0 (14; 1.0)
PG1_2_PG2 in [ "H" ] [ Pattern: 2 ] => 2.0 (0)
CA153 = H [ Pattern: 2 ] => 2.0 (13; 0.846)
```

图1 胃癌筛查决策树模型(规则集,部分)

Fig.1 Decision tree model for gastric cancer screening (rule set, partial)

2.3 胃癌筛查模型性能评估

基于预测的准确性,利用训练集和测试集分别对所建模型进行评估。结果发现模型对于训练集、测试集的准确率分别为89.58%、89.14%。根据测试集,利用受试者特征曲线(ROC)对模型进行评估,计算得到的曲线下面积(AUC)为0.809。

表5 对胃癌筛查重要性前10位的指标
Tab.5 Top 10 indicators for gastric cancer
screening ranked by importance

序号	指标	重要性
1	CA199	0.296 3
2	CA153	0.270 6
3	CEA	0.223 6
4	HCT	0.101 1
5	RDW_CV	0.031 0
6	Baso_Cnt	0.017 9
7	Lym_Per	0.011 2
8	PG1_2_PG2	0.009 9
9	PG2	0.009 0
10	DBIL	0.005 7

3 分析与讨论

3.1 胃癌筛查决策树模型的性能

利用临床业务数据,挖掘得到的胃癌筛查模型(图1)包含51条决策规则。评估结果显示,模型的整体准确率超过89%,是一套良好的决策规则集。

综合考虑模型筛查的灵敏度和特异性,从ROC曲线形状和曲线下面积(AUC=0.809)可以发现,所建模型具有良好的胃癌风险识别能力。

从规则的置信度来看,大多数规则具有较高的置信度,如以下规则的置信度为0.999:

CA153=N[Pattern:2](1, 718)

PG1_2_PG2 in [""L"] [Pattern:2]=> 2.0 (1, 601; 0.999)

由此可见,本研究所构建的胃癌筛查模型具有良好的分类性能和实用价值。

3.2 胃癌筛查指标的价值

从相关性分析(表3)、组间独立样本 t 检验(表4)以及决策树模型(图1)结果,可以发现一些关于胃癌筛查指标的有趣规律。

3.2.1 人口学因素 年龄和性别等人口学因素是胃癌的危险因素^[3]。胃癌组(A组)男性3 597例,女性1 966例,男女比例为1.83:1,与2008年IARC数据^[20]相近。性别与胃癌诊断具有显著相关性($P<0.05$),独立样本 t 检验结果也发现性别在胃癌组(A组)、非胃部恶性肿瘤组(B组)、非恶性肿瘤组(C组)之间均存在显著性差异(表4),性别在筛查模型中也是重要的决策属性之一。这些结果说明胃癌存在性别显著性差异,对于胃癌筛查具有一定价值,与已有报道^[4,20]

相一致。相关性分析结果(表3)显示,年龄与胃癌发病密切相关($P<0.05$),A组年龄明显高于C组,也高于B组,独立样本 t 检验也发现A、B、C组之间存在年龄显著性差异($P<0.05$)。说明年龄是胃癌的重要危险因素,其患病率随年龄增长而上升^[3]。然而,年龄在胃癌筛查模型中并未发现其作用。这可能说明,采用WHO方法对年龄进行划分,并不一定适合于我国胃癌筛查模型研究,这有待进一步探索。

3.2.2 血常规检查指标 谢燕等^[21]研究认为RDW、MPV、Lym_Cnt等血常规指标对于胃癌的早期诊断具有重要提示作用。本研究的相关性分析结果(表3)显示,部分血常规检查指标与胃癌诊断具有显著相关性,包括Baso_Per、Eos_Cnt、Hb、HCT、Lym_Cnt、Lym_Per、MCH、MCHC、MCV、Mono_Cnt、Mono_Per、MPV、Neu_Cnt、Neu_Per、P_LCR、PCT、PDW、PLT、RDW_CV、RDW_SD、WBC等。根据独立样本 t 检验结果(表4),发现Hb、HCT、Lym_Cnt、Lym_Per、MCH、MCHC、MCV、MPV、Neu_Cnt、Neu_Per、P_LCR、RDW_CV、TT、WBC在A、B、C3组之间均存在显著性差异。APTT、Baso_Per、Mono_Cnt、Mono_Per、PDW在非恶性肿瘤组(C组)与两个恶性肿瘤组(A组、B组)之间存在显著性差异。PCT、PLT在A组与B组、C组之间均存在显著性差异。Eos_Cnt、Eos_Per在B组与A组、C组之间均存在显著性差异。RBC仅在C组与B组之间存在显著性差异。Baso_Cnt仅在C组与A组之间存在显著性差异。在筛查模型中,HCT、Mono_Cnt、Baso_Cnt、PLT、Lym_Per、DBIL、APTT、MCH、Neu_Per等指标进入了决策属性。以上研究结果说明,血常规检查对于胃癌筛查具有一定价值。

3.2.3 血脂/肝功能及其它生化指标 相关性分析结果(表3)显示,血液生化检查中的ALB、ALB_2_GLB、DBIL、GLB、HDL_C、HDL_C_2_TC、LDH_L、LDL_C、TBA、TBIL、TC、TG、TP等与胃癌具有显著相关性。独立样本 t 检验结果发现ALB、ALB_2_GLB、DBIL、GLB、HDL_C、HDL_C_2_TC、TBA、TC、TG、TP在各组之间均存在显著性差异。LDH_L、LDL_C在非恶性肿瘤组(C组)与两个恶性肿瘤组(A组、B组)之间存在显著性差异,对于识别良恶性肿瘤具有一定价值。TBIL在A组与C组、B组之间均存在显著性差异,说明TBIL对于识别胃癌与非胃癌具有一定意义。

在胃癌筛查模型中,血液生化检查指标TP、DBIL、GLB进入了决策属性,对于胃癌的筛查具有重要作用。

3.2.4 肿瘤标志物指标 肿瘤标志物是肿瘤诊断的重要指标,主要的肿瘤标志物包括CA199、CA153、CA125、CEA、AFP、CA724等。本研究发现AFP、CA153、CA199与胃癌诊断具有显著相关性,而CEA、CA125未显示出显著相关性,如表3所示。独立样本 t 检验结果(表4)显示,AFP在A、B、C组3组之间均存在显著性差异,C组的CEA与A、B组之间均存在显著性差异,A组的CA125、CA199与C、B组均存在显著性差异,而B组与A组之间却不存在显著性差异,CA153仅在B组与A组之间具有显著性差异,而在C组与B、A组之间均无显著性差异。在胃癌筛查模型中,发现CEA、CA153、CA199均位列重要属性前3名(表5)。这些结果说明,肿瘤标志物CA199、CA153、CA125、CEA、AFP等在胃癌筛查中具有重要作用,但难以单独用于胃癌筛查。单一恶性肿瘤所特有的标志物至今尚未发现,联合检测更具诊断价值^[15,22]。通过联合使用以上肿瘤标志物,可以提高胃癌的鉴别能力,提高胃癌筛查的准确率。本研究未发现CA724在胃癌筛查中的价值,这与Wang等^[23]的研究结果一致。

3.2.5 其它指标 陈如昌等^[24]研究发现,FOB对胃癌的诊断和预后判断有重要的临床价值。本研究结果(表3)显示,FOB与胃癌诊断之间存在显著相关性。独立样本 t 检验结果(表4)也发现,FOB在各组之间均存在差异显著性,说明FOB对于胃癌的筛查具有重要价值。Cha等^[25]研究认为胃蛋白酶(PG)是胃部癌前病变和早期胃癌的有用标志物,文献^[20]也认为PG检测是一种有用的胃癌检查标志物。本研究相关性分析(表3)、独立样本 t 检验(表4)、决策树模型(图1)均提示PG1_2_PG2在胃癌筛查中具有重要作用。可见,胃蛋白酶检测可作为早期胃癌筛查的手段之一。从相关性分析结果看, Hp 与胃癌诊断有显著相关性,C组与A、B组之间存在显著性差异,但A组与B组之间无显著性差异,也未进入辅助筛查模型的决策变量中,说明 Hp 对于胃癌筛查作用不充分,与文献^[26]报道相一致。

4 结论

综上所述,对于胃癌筛查,实验室检查是一种费用低廉、方便有用的工具。有价值的检测指标包括RH、AFP、ALB、ALB_2_GLB、APTT、Baso_Cnt、Baso_Per、CA125、CA153、CA199、CEA、DBIL、D_Dimer、Eos_Cnt、Eos_Per、FOB、GLB、Hb、HCT、HDL_C、HDL_C_2_TC、 Hp 、LDH_L、LDL_C、Lym_Cnt、Lym_Per、MCH、MCHC、MCV、Mono_Cnt、

Mono_Per、MPV、Neu_Cnt、Neu_Per、P_LCR、PCT、PDW、PG1、PG1_2_PG2、PG2、PLT、PT、RBC、RDW_CV、TBA、TBIL、TC、TG、TP、TT、WBC等。其中,CA199、CA153、CEA、HCT、RDW_CV、Baso_Cnt、Lym_Per、PG1_2_PG2、PG2、DBIL为胃癌筛查的前10个重要指标,胃癌筛查可重点考虑使用这些检验项目。利用临床数据,通过决策树算法,可以建立性能良好的胃癌筛查模型,对于提高胃癌筛查水平具有参考价值。

【参考文献】

- [1] World Life Expectancy. USA causes of death by age and gender[EB/OL]. <https://www.Worldlifeexpectancy.com/usa-cause-of-death-by-age-and-gender>, 2018-10-9.
- [2] CHEN W, ZHENG R, BAAD P D, et al. Cancer statistics in China, 2015[J]. CA Cancer J Clin, 2016, 6(2): 115-132.
- [3] 廖专, 孙涛, 吴浩, 等. 中国早期胃癌筛查及内镜诊治共识意见(2014年4月·长沙)[J]. 胃肠病学, 2014, 19(7): 408-427.
LIAO Z, SUN T, WU H, et al. Chinese expert consensus on screening and endoscopic management of early gastric cancer (Changsha, 2014)[J]. Chinese Journal of Gastroenterology, 2014, 19(7): 408-427.
- [4] 郑朝旭, 郑荣寿, 张思维, 等. 中国2010年胃癌发病与死亡分析[J]. 中国肿瘤, 2014, 23(10): 795-800.
ZHENG Z X, ZHENG R S, ZHANG S W, et al. An analysis of incidence and mortality of stomach cancer in China, 2010[J]. China Cancer, 2014, 23(10): 795-800.
- [5] GE S, FENG X, SHEN L, et al. Association between habitual dietary salt intake and risk of gastric cancer: a systematic review of observational studies[J]. Gastroent Res Pract, 2012(5): 808120.
- [6] ESLICK G D, LIM L L, BYLES J E, et al. Association of *Helicobacter pylori* infection with gastric carcinoma: a meta-analysis[J]. Am J Gastroenterol, 1999, 94(9): 2373-2379.
- [7] 刘文忠, 谢勇, 陆红, 等. 第五次全国幽门螺杆菌感染处理共识报告[J]. 中国实用内科杂志, 2017, 37(6): 509-524.
LIU W Z, XIE Y, LU H, et al. Fifth Chinese national consensus report on the management of *Helicobacter pylori* infection[J]. Chinese Journal of Practical Internal Medicine, 2017, 37(6): 509-524.
- [8] KLUIJT I, SIJMONS R H, HOGERBRUGGE N, et al. Dutch Working Group on Hereditary Gastric Cancer. Familial gastric cancer: guidelines for diagnosis, treatment and periodic surveillance[J]. Fam Cancer, 2012, 11(3): 363-369.
- [9] 赵凤源, 贺圣文. 胃癌危险因素研究进展[J]. 现代预防医学, 2010, 37(11): 2105-2106.
ZHAO F Y, HE S W. Research progress on risk factors of gastric cancer[J]. Modern Preventive Medicine, 2010, 37(11): 2105-2106.
- [10] TSENG C H, TSENG F H. Diabetes and gastric cancer: the potential links[J]. World J Gastroenterol, 2014, 20(7): 1701-1711.
- [11] CROSBY D A, DONOHOE C L, FITZGERALD L, et al. Gastric neuroendocrine tumours[J]. Dig Surg, 2012, 29(4): 331-348.
- [12] ISOBE Y, NASHIMOTO A, AKAZAWA K, et al. Gastric cancer treatment in Japan: 2008 annual report of the JGCA nationwide registry[J]. Gastric Cancer, 2011, 14(4): 301-316.
- [13] AJANI J A, BENTREM D J, BESH S, et al. Gastric cancer, version 2. 2013: featured updates to the NCCN guidelines[J]. J Natl Compr Canc Netw, 2013, 11(5): 531-546.
- [14] MIKI K, MORITA M, SASAJIMA M, et al. Usefulness of gastric

- cancer screening using the serum pepsinogen test method[J]. Am J Gastroenterol, 2003, 98(4): 735-739.
- [15] 周萍, 王聪, 赵车冬, 等. CEA、CA199、CA125、CA724联合检测对胃癌的临床诊断价值[J]. 实用癌症杂志, 2017, 32(9): 1427-1429. ZHOU P, WANG C, ZHAO C D, et al. The clinical application value of combined detection of tumor markers CEA, CA199, CA125 and CA724 in gastric cancer[J]. The Practical Journal of Cancer, 2017, 32(9): 1427-1429.
- [16] 陈西. 化学法粪便隐血试验在食管癌、胃癌患者出血评估中的应用[J]. 中国医药指南, 2017, 15(34): 8-9. CHEN X. Application of chemical fecal occult blood test in assessment of bleeding in patients with esophageal cancer and gastric cancer[J]. Guide of China Medicine, 2017, 15(34): 8-9.
- [17] 李康, 旦增, 巴桑, 等. 粪隐血试验在高原地区社区人群胃癌初筛中的效能探讨[J]. 中国慢性病预防与控制, 2013, 21(5): 601-602. LI K, DAN Z, BA S, et al. Effectiveness of fecal occult blood test for screening gastric cancer among community population in plateau area[J]. Chinese Journal of Prevention and Control of Chronic Diseases, 2013, 21(5): 601-602.
- [18] 王云玲, 宋娟, 杜江华, 等. 双源CT能谱成像定性评估胃癌转移淋巴结[J]. 中国医学影像技术, 2018, 34(5): 705-708. WANG Y L, SONG J, DU J H, et al. Dual-source CT energy spectral imaging in qualitative evaluation of metastatic lymph nodes in patients with gastric cancer[J]. Chinese Journal of Medical Imaging Technology, 2018, 34(5): 705-708.
- [19] XU Z Q, BROZA Y Y, IONSECU R, et al. A nanomaterial-based breath test for distinguishing gastric cancer from benign gastric conditions[J]. Br J Cancer, 2013, 108(4): 941-950.
- [20] FERLAY J, SHIN H R, BRAY F, et al. GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide: IARC CancerBase No.10 [J]. Int J Cancer, 2010, 136(5): E359-E386.
- [21] 谢燕, 文剑波, 李兴. 血常规分析对胃癌的早期诊断价值及与幽门螺旋杆菌感染的相关性[J]. 世界华人消化杂志, 2018, 26(15): 904-911. XIE Y, WEN J B, LI X. Value of routine blood test in early diagnosis of gastric cancer and its relationship with *Helicobacter pylori* infection[J]. World Chinese Journal of Digestology, 2018, 26(15): 904-911.
- [22] 杨丽萍, 田家强. 肿瘤标志物CA19-9、AFP、CEA及CA72-4联合诊断老年消化系统恶性肿瘤的价值研究[J]. 国际检验医学杂志, 2018, 39(16): 2026-2029. YANG L P, TIAN J Q. The value of tumor markers CA19-9, AFP, CEA and CA72-4 in the diagnosis of senile digestive system malignancies[J]. International Journal of Laboratory Medicine, 2018, 39(16): 2026-2029.
- [23] WANG S L, YU G Y, YAO J, et al. Diagnostic role of carbohydrate antigen 72-4 for gastrointestinal malignancy screening in Chinese patients: a prospective study[J]. J Dig Dis, 2018, 19(11): 685-692.
- [24] 陈如昌, 陈俊, 吴健媛. 粪便隐血和肿瘤标志物联合检验在胃癌诊治中的价值研究[J]. 中国卫生检验杂志, 2015, 25(22): 3868-3870. CHEN R C, CHEN J, WU J Y. Research on the clinical value of combined detection of FOB and CEA in the diagnosis and treatment of gastric cancer[J]. Chinese Journal of Health Laboratory Technology, 2015, 25(22): 3868-3870.
- [25] CHA J H, JANG J S. Clinical correlation between serum pepsinogen level and gastric atrophy in gastric neoplasm[J]. Korean J Intern Med, 2018. doi: 10.3904/kjim.2018.282.
- [26] CHISATO H. Update version of the Japanese guidelines for gastric cancer screening[J]. Jpn J Clin Oncol, 2018, 48(7): 673-683.

(编辑:黄开颜)