基于 Logistic 回归的食管癌风险预测模型研究

曹云珍 高皓楠

（指导教师：张艳萍）

河北工程大学

一、研究背景

食管癌是世界上最常见的六大恶性肿瘤之一，也是中国最常见的恶性肿瘤，居中国恶性肿瘤死因的第四位。它是一种常见的消化道恶性肿瘤，全世界每年新发病例约 30 万。食管癌起病隐匿，大多数患者就诊时已经是中晚期，5 年生存率仅有 20%。对于早期的食管癌患者，目前临床上主要采用以手术为主的根治性治疗方式，而对于中晚期无法手术的患者，则主要采用化学治疗和放射治疗。近年来，尽管食管癌的治疗技术已经有了很大的进步，但食管癌恶性程度高，病程进展迅速，易复发和转移，患者的总体预后仍然比较差。因此，建立在临床上有应用前景的食管癌风险预测模型，对于筛查食管癌高危人群，实现食管癌患者的早诊早治疗，改善食管癌患者的生存情况具有重要的意义。

通过预测模型可以确定患者未来发病风险。疾病风险预测模型不仅在发病率较高的高血压等[1]应用广泛，在肿瘤发病率相对较低的食管癌等[2]也有应用。食管癌的病理类型在不同国家地区有所差异。在西方国家，食管癌的病理类型以食管腺癌为主，且与 Barrett 食管癌的发病密切相关。而在亚洲国家，食管癌以食管鳞状细胞癌（ESCC）为主，占到了全部食管癌的 90%以上[3]，我国河北涉县、磁县等地是食管癌高发地区[4]。近年来国内外研究者对食管癌进行了大量的流行病学研究和病因学研究，从不良生活方式和饮食习惯等多方面进行了探索，取得了有意义的进展，为食管癌的防治提供了一定的科学依据。为了更好的为邯郸市食管癌高危人群预警、早期诊断、个体化防治提供更有力的理论依据，本文探索影响邯郸市居民食管癌发病的主要因素，并建立其食管癌风险预测模型[5-9]。

二、研究目的

通过收集邯郸市某医院 2017 年食管癌住院患者信息和居民健康人群调查问

卷信息，整理影响食管癌患者患病的相关数据，对患者性别、年龄、居住地、吸烟情况、饮酒情况、个人健康情况、从事工作、是否患食管癌进行简单描述性统计和相关性分析，并建立 Logistic 风险预测模型，以期达到以下目的：

1. 了解食管癌患者的基本特点；
2. 找到影响食管癌患者患病的主要因素；
3. 通过建立的风险预测模型，对患者进行风险等级评价。

三、数据来源

通过收集邯郸市某医院 2017 年食管癌住院患者信息和居民健康人群调查问卷信息，最后分别得到 100 个样本和 114 个样本。我们定义抽烟者、饮酒者的标准如下：平均每周至少两次并达一年的人定义为饮酒者，否则为非饮酒者；一生种吸烟总量大于 100 支或吸烟斗大于 100 次的人定义为吸烟者，否则为非吸烟者。因变量为是否患有食管癌，自变量为性别（X_1）、年龄（X_2）、居住地（X_3）、吸烟情况（X_4）、饮酒情况（X_5）、个人健康情况（X_6）、从事工作（X_7）。

表 1 变量说明

变量	说明
性别 X_1	0=男 1=女
年龄 X_2	0=<18 1=18-25 2=26-30 3=31-40 4=41-50 5=51-60 6=>60
居住地 X_3	0=邯山区 1=丛台区 2=复兴区 3=其它
吸烟情况 X_4	0=从不 1=吸烟
饮酒情况 X_5	0=从不 1=喝酒
个人健康情况 X_6	0=高血压 1=糖尿病 2=冠心病 3=脑梗塞 4=胃病 5=其它
从事工作 X_7	0=退休 1=工作
是否患食管癌	0=否 1=是

四、描述统计

（一）Spearman 相关性分析

Spearman 相关系数被定义成等级变量之间的皮尔逊相关系数。本文的因变量和自变量均为等级离散型变量，因此采用 Spearman 相关系数对变量之间的相关性及性别、年龄、居住地、吸烟情况、饮酒情况、个人健康情况、从事工作与是否患食管癌之间的相关关系进行分析，结果如表 2 所示。

表 2　变量之间的 spearman 相关性分析

		性别	年龄	居住地	吸烟情况	饮酒情况	个人健康情况	从事工作	是否患食管癌
性别	相关系数	1.000	.098	.040	-.437**	-.391**	-.029	-.016	.042
	显著性	.	.153	.559	.000	.000	.677	.812	.546
年龄	相关系数	.098	1.000	.398**	-.027	-.361**	-.386**	-.113	.519**
	显著性	.153	.	.000	.691	.000	.000	.101	.000
居住地	相关系数	.040	.398**	1.000	-.081	-.322**	-.017	.271**	.661**
	显著性	.559	.000	.	.238	.000	.806	.000	.000
吸烟情况	相关系数	-.437**	-.027	-.081	1.000	.521**	-.092	-.198**	-.134*
	显著性	.000	.691	.238	.	.000	.181	.004	.050
饮酒情况	相关系数	-.391**	-.361**	-.322**	.521**	1.000	.004	-.269**	-.480**
	显著性	.000	.000	.000	.000	.	.953	.000	.000
个人健康情况	相关系数	-.029	-.386**	-.017	-.092	.004	1.000	.159*	-.101
	显著性	.677	.000	.806	.181	.953	.	.020	.141
从事工作	相关系数	-.016	-.113	.271**	-.198**	-.269**	.159*	1.000	.530**
	显著性	.812	.101	.000	.004	.000	.020	.	.000
是否患食管癌	相关系数	.042	.519**	.661**	-.134*	-.480**	-.101	.530**	1.000
	显著性	.546	.000	.000	.050	.000	.141	.000	.

**. 在 0.01 级别（双尾），相关性显著。

*. 在 0.05 级别（双尾），相关性显著。

表 2 结果显示自变量之间存在着一定的线性相关性，如性别与吸烟情况、饮酒情况（P=0.000<0.01），而与是否患食管癌的相关性变量中，性别和个人健康情况显然与是否患食管癌不具有显著相关性，其 P 值均大于 0.05，表示在 0.05 水平上两者之间不具有显著相关性。吸烟情况与是否患食管癌的 P=0.05，在 0.05

水平上不易判断，因此进一步进行列联表分析和卡方分析。

在其它变量的相关分析中，饮酒情况与患食管癌具有负相关关系（$\rho < 0$），年龄、居住地、从事工作与是否患食管癌具有正相关关系（$\rho > 0$）。

（二）交叉列联表分析和卡方检验

表 3 自变量与因变量交叉列联分析和卡方检验

		是否患食管癌		卡方	P
		否	是		
性别	男	73	60	0.369	0.544
	女	41	40		
年龄	小于 18	1	0	65.148	0.000
	18-25	12	0		
	26-30	2	0		
	31-40	17	0		
	41-50	28	3		
	51-60	14	17		
	大于 60	40	80		
居住地	邯山区	58	4	95.643	0.000
	丛台区	19	5		
	复兴区	15	6		
	其它	22	85		
吸烟情况	从不	65	70	3.855	0.050
	吸烟	49	30		
饮酒情况	从不	46	87	49.280	0.000
	饮酒	68	13		
个人健康情况	高血压	31	27	16.013	0.007
	糖尿病	4	4		
	冠心病	0	4		
	脑梗塞	4	1		
	胃病	9	22		
	其它	66	42		
从事工作	退休	60	4	60.101	0.000
	工作	54	96		

分别做性别、年龄、居住地、吸烟情况、饮酒情况、个人健康情况、从事工作与是否患食管癌之间的交叉列联表，并对其进行卡方检验，所有结果综合在表3所示，与Spearman相关性（表2）分析结果相似，性别与是否患食管癌不具有显著相关性（χ^2=0.369，P=0.544>0.05）；年龄对是否患食管癌具有显著性影响（χ^2=65.148，P=0.000<0.05），可以发现食管癌患者年龄普遍较高，即年龄越大，食管癌患病风险越高；食管癌患病情况在四个居住地分布情况上具有显著差异（χ^2=95.643，P=0.000<0.05），未患食管癌被调查者在四个区域分布均匀，但患病人员集中分布在其它区，考虑这与城镇化具有相关关系，城市内的人员定期进行身体检查，对于防范食管癌患病，早发现早治疗具有一定的促进作用；饮酒情况对是否患食管癌具有显著相关性（χ^2=49.280，P=0.000<0.05），患病人员饮酒情况较少，说明合理饮酒对于防患食管癌具有一定的效果；从事工作情况与是否患食管癌具有显著相关性（χ^2=60.101，P=0.000<0.05），食管癌患者更多是从事相关的工作。

通过自变量与因变量的列联表分析以及 Spearman 相关系数矩阵得到年龄（X_2）、居住地（X_3）、吸烟情况（X_4）、饮酒情况（X_5）、从事工作类型（X_7）与患有食道癌具有显著的相关性。而性别（X_1）、个人健康情况（X_6）与患有食管癌没有显著的相关性。

五、食管癌风险预测模型的构建

（一）Logistic 回归模型建立

从相关分析可以看出，性别（X_1）、个人健康情况（X_6）与因变量没有统计学相关性。因此 Logistic 回归模型中，纳入以下自变量：年龄（X_2）、居住地（X_3）、吸烟情况（X_4）、饮酒情况（X_5）、从事工作（X_7）。从而建立 $\log it(p)$ 关于自变量 X_2, X_3, X_5, X_7 的 Logistic 回归方程为：

$$\log it(p) = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_5 + \beta_4 X_7$$

利用 R 语言进行数据处理和分析，各回归系数中吸烟情况（X_4）的 P>0.05，从而得到此变量对 Logistic 回归模型没有显著的统计学意义。进一步，利用逐步

回归方法对模型进行优化，发现去除吸烟情况（X_4）变量后，所有的回归系数 P <0.05，说明各变量对 Logistic 回归优化模型具有显著的统计学意义。

为了验证优化模型的有效程度，对原模型和优化模型的卡方检验进行了比较，如表 4 所示：

表 4　原模型和优化模型比较

模型	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
原模型	208	15.380			
优化模型	209	15.504	-1	-0.12438	0.1946

从表中得到，卡方值 $P=0.1946>0.05$，得到吸烟情况（X_4）这个变量不会显著影响模型的预测精度，从而验证了优化模型的有效程度。

最终得到是否患有食管癌的 Logistic 回归优化模型为：

$$\log it(p)=\ln\left(\frac{p}{1-p}\right)=-0.72414+0.13026X_2+0.12750X_3-0.11173X_5+0.50541X_7$$

（二）模型验证及指标分析

为了避免引入过多变量导致模型的过度拟合，以至于预测的严重失真，通过 5 折交叉验证方法进行检验，得到平均误差为 7.01%，说明模型基本不存在过拟合现象。模型预测准确度为 92.99%，并通过最优模型 Logistic 回归中的受试者工作特征曲线(receiver operating characteristic curve, ROC 曲线)，计算了相应的曲线下面积(areas under the curve, AUC)为 0.985（如图 1 所示），从而说明建立的 Logistic 回归优化模型是高度有效的。

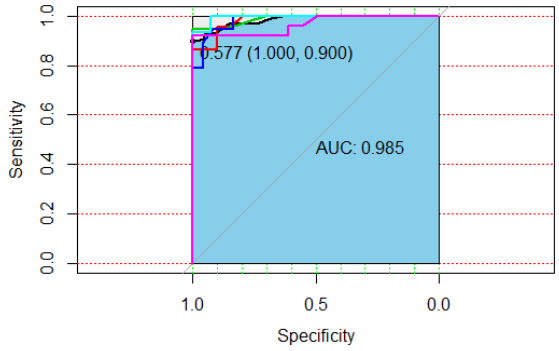

图 1　ROC 曲线及 AUC 值

进一步，对回归系数的风险比值比（odds ratio, OR）进行了分析，随着年龄的变化，每增大一个单位，患有食管癌的概率就会增加 1.139 倍；随着居住地的变化，每变化一个单位，患有食道癌的概率就增加 1.136 倍；随着饮酒情况的变化，饮酒者是非饮酒者患有食道癌概率的 0.894 倍；工作人群患有食道癌的概率是退休人群的 1.658 倍。

综上所述，邯郸市居民食管癌的患病率主要跟其生活习惯以及居住环境的工作人群相关，但是食管癌相关危险因素的影响仍需进一步研究和探讨，从而建立准确性更高的食管癌风险预测模型，为食管癌高危人群预警、早期诊断、个体化防治提供更有力的理论依据。

六、研究结论

本文首先通过 Spearman 相关性分析和卡方列联表分析对食管癌患者进行初步分析，结果发现年龄、居住地、吸烟情况、饮酒情况、从事工作类型与患有食道癌具有显著的相关性。而性别、个人健康情况与患有食管癌没有显著的相关性。具体来说，性别与是否患食管癌不具有显著相关性；年龄对是否患食管癌具有显著性影响，食管癌患者年龄普遍较高，即年龄越大，食管癌患病风险越高；食管癌患病情况在四个居住地分布情况上具有显著差异，未患食管癌被调查者在四个区域分布均匀，但患病人员集中分布在其它区，考虑这与城镇化具有相关关系，城市内的人员定期进行身体检查，对于防范食管癌患病，早发现早治疗具有一定的促进作用；饮酒情况对是否患食管癌具有显著相关性，患病人员饮酒情况较少，说明合理饮酒对于防患食管癌具有一定的效果；从事工作情况与是否患食管癌具有显著相关性，食管癌患者更多是从事相关的工作。

其次基于 Logistic 回归模型建立了最优食管癌的风险预测模型。影响邯郸市居民食管癌发病的主要因素为年龄、居住地、饮酒情况、从事工作。并通过逐步回归方法得到最优的 Logistic 回归预测模型，同时应用 5 折交叉验证方法验证了模型基本不存在过拟合现象，预测准确度达到 92.99%。食管癌风险预测模型的建立为邯郸市食管癌高危人群预警、早期诊断、个体化防治可提供更有力的理论依据。

七、建议

1、定期到医院检查

首先随着年龄的增长，食管癌患病风险逐步增高，因此建议家里老人定期到医院进行相关检查，在食管癌预防方面，应进行胃镜检查和消化道造影检查，通过观察病变组织、病变部位等来确诊，及时获得医生帮助，进行早发现、早治疗，减少食管癌患者病死率。

其次，在相关性分析中，我们发现居住地不同对患食管癌概率不同，不同的居住地生活环境、医疗条件都不同。据了解，邯山区、复兴区、丛台区医疗条件较其它区域都更加完善，建有健康小屋等专门服务于社区人员慢性疾病的医疗机构，会为社区人员进行定期检查，并进行跟踪，相比之下涉县等地进行检查则需到医院进行就诊，较为繁琐，因此医疗设施的健全对于早期发现食管癌患者具有重要的影响作用，因此建议居住在离城市较为偏远地区的人群也应提高防范意识，定期到医院进行检查，并建议相关部门建立相关辅助机构，可定期下乡，以在早期发现食管癌患者，减少食管癌患者病死率。

2、合理饮食，注意生活习惯

在食管癌患者的基本情况分析中，饮酒情况、吸烟情况以及工作情况对是否患食管癌具有一定的影响，结合临床经验，不仅仅是食管癌患者，每个人都应注意饮酒、吸烟以及饮食等问题，对于吸烟人员应采取一定措施减少吸烟或进行戒烟，对于饮酒人员应注意适当饮酒，而在饮食方面应注意养成吃饭不挑食，少吃或不吃反季蔬菜，注意荤素搭配，多吃粗粮，每餐七分饱即可，在吃饭时应注意细嚼慢咽多吃温食，忌凉食、热食，少吃或不吃咸菜，不吃霉变食物。其次，关于从事工作方面，食管癌患者应注意劳逸结合，关注自己身体，量力而行，避免过度劳累加剧疾病恶化。

参考文献

[1] Cardiovascular risk. Ceneva, World Health Organization, 2007.

[2] 李婧, 秦江梅. 新疆哈萨克族食管癌风险的预测 [J], 世界华人消化杂志, 2014, 22(10): 1442-1445.

[3] Jemal A, Siegel R, Ward E, et al. Cancer statistics [J], CA Cancer J Clin, 2009, 59(4) : 225-249.

[4] 赵志敏, 李秀敏, 贺晓. 豫北农村社区食管癌贲门癌患者的生存质量及影响因素 [J]. 中国全科医学, 2010, 13 (19): 2130-2133.

[5] 孙振球, 徐勇勇. 医学统计学 [M]. 人民卫生出版社, 2010, 278-336.

[6] 李丽霞, 王彤. BP 神经网络与 Logistic 回归的比较研究 [J]. 中国卫生统计, 2005, 2(3): 138-140.

[7] 李运明, 徐勇勇. 国人健康风险模型及风险评估方法研究 [M]. 第四军医大学博士学位论文，2011: 70-103.

[8] 张家放. 医用多元统计方法 [M]. 高等教育出版社, 2004, 128-157.

[9] 章杨熙. 医学统计预测 [M]. 中国科学技术出版社, 1995, 90-96.

附录：本案例所使用 R 软件程序命令

```
rm(list = ls()) #  清空变量
install.packages("psych")
library(psych)
library(pROC)
data = read.csv("F:/学校学习/报告/食管癌/data.csv",header = T)#读取数据
cor(data,method = "spearman")#spearman 相关性分析
corr.test(data,method = "spearman")#相关系数检验
table(data$性别,data$食道癌)#列联表分析
p_value = rep(1,length(data)-1)
for(i in 1:(length(data)-1)){
p = chisq.test(x = data[i],y = data$食道癌,correct = T,rescale.p = F)
p_value[i] = p$p.value
}
p_value#卡方分析
model0 = glm(食道癌~年龄+居住地+吸烟情况+饮酒情况+从事工作,family =
binomial(link=logit),data = data)
summary(model0)#初步建立模型
model1 <- step(model0,direction="both")#逐步筛选法变量选择
summary(model1)#最优模型相关信息
p = predict(model1,data[1:7])#利用模型 model1 对数据 data 进行预测
p = exp(p)/(1+exp(p))#计算预测得到的概率
ngrids=100#设置格点数为 100
TPR=rep(0,ngrids)#为 TPR(true positive ratio)赋初值
FPR=rep(0,ngrids)#为 FPR(false positive ratio)赋初值
for(i in 1:ngrids){
p0 = i/ngrids;#选取阈值 p0
y.true = data$食道癌#从 a2 中取出真实值并赋值给 y.true
y.pred = 1*(p>p0)#以 0.05 为阈值生成预测值
```

```
TPR[i] = sum(y.pred*y.true)/sum(y.true)#计算 TPR
FPR[i] = sum(y.pred*(1-y.true))/sum(1-y.true)#计算 FPR
}
data_roc = roc(y.true,p)#画 ROC 曲线
plot(data_roc)
data_AUC = auc(data_roc)#计算 AUC 值
data_AUC
```