四川大学学报（医学版） 2023，54（5）：892－898
J Sichuan Univ（Med Sci）

在线全文

大规模人群队列生活行为方式相关的肺癌风险预测模型的构建*

陈睿琳[1]，王静茹[1]，王 硕[1]，唐思琦[2]，索 晨[1,2,3△]

1. 复旦大学公共卫生学院 流行病学教研室（上海 200032）；2. 上海市重大传染病和生物安全研究院（上海 200032）；
3. 复旦大学泰州健康科学研究院（泰州 225316）

【摘要】 目的 发现影响肺癌发病的生活行为相关危险因素，并构建肺癌风险预测模型，识别人群中的高风险个体，帮助肺癌早期筛查。**方法** 本研究数据来源于英国生物样本库（UK Biobank）2006年3月–2010年10月收集的502 389名参与者。参考国内外肺癌筛查指南和高质量肺癌危险因素研究文献，确定本研究高危人群识别标准。采用单因素Cox回归分析及逐步回归筛选出肺癌的危险因素，通过Cox比例风险回归构建多因素肺癌风险预测模型，根据比较赤池信息准则以及Schoenfeld残差检验结果，最终选择等比例假设的最优拟合模型。多因素Cox比例风险回归考虑生存时间，将人群按7∶3的比例随机分为训练集和验证集，使用训练集建立模型，并用验证集对模型性能进行内部验证。受试者工作特征曲线（ROC）曲线的曲线下面积（*AUC*）被用于评估模型的效能。将人群按照发病概率的0%～<25%、25%～<75%、75%～100%分为低风险、中风险及高风险人群，分别计算其中的发病人数占比。**结果** 本研究最终纳入453 558人，在累计随访5 505 402人年期间，共诊断出2 330例肺癌。Cox比例风险回归分析筛选出10个自变量建立模型：年龄、体质量指数（body mass index, BMI）、学历、收入、体力活动情况、吸烟状态、饮酒频率、新鲜水果摄入量、癌症家族史、烟草暴露。该模型通过内部验证结果显示8个自变量（除BMI和新鲜水果摄入量外）均是肺癌的影响因素（P<0.05）。该模型训练集预测肺癌发生的一年、五年、十年*AUC*分别为0.825、0.785、0.777；验证集预测肺癌发生的一年、五年、十年*AUC*分别为0.857、0.782、0.765。筛查高风险人群可发现68.38%的未来肺癌发病个体。**结论** 本研究建立了大规模人群生活行为方式相关的肺癌风险预测模型，其在判别能力方面表现出良好的性能，为制定肺癌标准化筛查策略提供了工具。

【关键词】 肺癌 风险预测 预测模型 危险因素

Construction of a Risk Prediction Model for Lung Cancer Based on Lifestyle Behaviors in the UK Biobank Large-Scale Population Cohort

CHEN Ruilin[1], *WANG Jingru*[1], *WANG Shuo*[1], *TANG Siqi*[2], *SUO Chen*[1,2,3△]. 1. *Department of Epidemiology, School of Public Health and Key Laboratory of Public Health Safety of Ministry of Education, Fudan University*, Shanghai 200032, China; 2. *Shanghai Institute of Infectious Disease and Biosecurity, Fudan University*, Shanghai 200032, China; 3. *Fudan University Taizhou Institute of Health Sciences*, Taizhou 225316, China

△ Corresponding author, E-mail: suochen@fudan.edu.cn

【Abstract】 Objective To identify the risk factors related to lifestyle behaviors that affect the incidence of lung cancer, to build a lung cancer risk prediction model to identify, in the population, individuals who are at high risk, and to facilitate the early detection of lung cancer. **Methods** The data used in the study were obtained from the UK Biobank, a database that contains information collected from 502 389 participants between March 2006 and October 2010. Based on domestic and international guidelines for lung cancer screening and high-quality research literature on lung cancer risk factors, high-risk population identification criteria were determined. Univariate Cox regression was performed to screen for risk factors of lung cancer and a multifactor lung cancer risk prediction model was constructed using Cox proportional hazards regression. Based on the comparison of Akaike information criterion and Schoenfeld residual test results, the optimal fitted model assuming proportional hazards was selected. The multiple factor Cox proportional hazards regression was performed to consider the survival time and the population was randomly divided into a training set and a validation set by a ratio of 7:3. The model was built using the training set and the performance of the model was internally validated using the validation set. The area under the receiver operating characteristic (ROC) curve (*AUC*) was used to evaluate the efficacy of the model. The population was categorized into low-risk, moderate-risk, and high-risk groups based on the probability of occurrence of 0% to <25%, 25% to <75%, and 75% to 100%. The respective proportions of affected individuals in each risk group were calculated. **Results** The study eventually covered 453 558 individuals, and out of the cumulative follow-up of 5 505 402 person-years, a total of 2 330 cases of lung cancer were diagnosed. Cox proportional hazards regression was performed to identify 10 independent variables as predictors of lung cancer, including age, body mass index (BMI), education, income, physical activity, smoking status, alcohol consumption frequency, fresh fruit intake, family history of cancer, and tobacco exposure, and a model was established accordingly.

* 国家重点研发计划项目（No. 2022YFC3400700、No. 2019YFC1315804），上海市公共卫生体系建设三年行动计划优秀人才项目（No. GWV-10.2-YQ32），上海市市级科技重大专项（No. ZD2021CY001）和上海科技委员会创新基金（No. 20ZR1405600）资助

△ 通信作者，E-mail：suochen@fudan.edu.cn

Internal validation results showed that 8 independent variables (all the 10 independent variables screened out except for BMI and fresh fruit intake) were significant influencing factors of lung cancer (P<0.05). The *AUC* of the training set for predicting lung cancer occurrence at one year, five years, and ten years were 0.825, 0.785, and 0.777, respectively. The *AUC* of the validation set for predicting lung cancer occurrence at one year, five years, and ten years were 0.857, 0.782, and 0.765, respectively. 68.38% of the individuals who might develop lung cancer in the future could be identified by screening the high-risk population. **Conclusion**　We established, in this study, a model for predicting lung cancer risks associated with lifestyle behaviors of a large population. Showing good performance in discriminatory ability, the model can be used as a tool for developing standardized screening strategies for lung cancer.

【Key words】　Lung cancer　Risk prediction　Prediction model　Risk factor

肺癌是世界主要高发恶性肿瘤之一，也是全世界癌症死亡谱中的重要组成部分。最新全球癌症数据表明，2020年新发肺癌病例约221万例（占比11.4%），死亡约180万例（占比18.0%）[1]。我国癌症中心登记数据显示，2016年我国肺癌新发病例约82.81万，粗发病率为59.89/10^5；肺癌死亡约65.70万例，死亡率为47.51/10^5，均位居所有恶性肿瘤发病谱和死亡谱的首位。随着我国人口老龄化加剧、工业化和城镇化进程的推进，不健康生活方式、环境暴露等危险因素的累加，未来的恶性肿瘤防控形势依旧十分严峻[2]。

根据《中国肺癌筛查与早诊早治指南》的建议，高危人群且年龄处于50～74岁者，应定期进行低剂量 CT检查。虽然在过去的数十年中，肺癌的诊断和治疗水平已经得到了极大的提高，但是预后改善效果并不显著。在我国，2003–2015年，肺癌的5年生存率虽有轻微提升，但是仍然低于20.0%；2012–2015年，中国肺癌患者的5年生存率只有19.7%，而决定其5年生存率的主要因素是其临床确诊时的肿瘤分期。临床资料表明，早期肺癌的5年生存率接近61.2%，其中小于1 cm的Ⅰ期肺癌5年生存率更是高达92%；然而，晚期肺癌的5年生存率却只有7.0%[3]。因此，早期诊断是提高肺癌预后的关键，然而关于最佳筛查策略（例如最佳筛查间隔）的争论仍在进行。国家癌症中心始终积极推动施行恶性肿瘤防治部署，完善恶性肿瘤的早发现、早诊断、早治疗。但是，由于不同地区的医疗水平、经济状况差异明显，且没有具体的肺癌检测技术标准，限制了对肺癌早期诊断和治疗的效果和效益。

普及戒烟限酒、合理膳食、适量运动和心理平衡等健康生活方式，提高群众自我防控意识和能力，是恶性肿瘤防治的重要措施[4]。近年来，国内外学者以不同特征人群为基础，陆续构建出了多种肺癌风险预测模型，既往的相关研究强调了一个事实，即与生活方式相关的环境危险因素可以为肺癌的风险预测提供策略，以便识别出高风险个体，在疾病早期即控制这些环境因子，或减缓其进展，从而降低肺癌疾病负担。

针对以上情况，本研究在开展临床检查前对肺癌发病的生活行为相关危险因素进行分析，构建肺癌发病风险预测模型，从而识别出可能受益于筛查策略的高危个体，在疾病早期对其进行预筛查，提高早期肺癌的检出率，并及时采取干预措施，从而降低肺癌疾病负担。

1　对象与方法

1.1　研究对象

本研究数据来源于英国生物样本库（UK Biobank, UKB）大规模前瞻性队列（申请编号为92718），该队列于2006–2010年期间招募502 389名英国居民，参与者进入队列时的年龄为40～69岁，其中3人在前往研究中心的当天死亡。该数据库的使用获得North West-Haydock Research Ethics Committee的批准（IRAS project ID: 299116）。

基于来自英国国家肿瘤登记处的肿瘤登记数据，本研究排除了进入队列前即确诊肿瘤的对象46 385人及协变量数据缺失的参与者2 446人后，共453 558人被纳入最终分析。研究起点定义为参与者进入队列的日期，随访时间从进入队列6个月后开始计算，至确诊肺癌、失访、死亡、退出研究或完成随访为止，截止日期至2022年11月30日。

根据UKB的英国国家肿瘤登记数据，使用国际疾病分类第十版（International Classification of Diseases, Tenth Revision, ICD-10）对肿瘤进行编码，本研究肺癌患者的定义为诊断ICD编码为肺癌，即ICD-10: C34的住院患者。

1.2　危险因素筛选

UKB的所有参与者均由在英国专门设置的22个评估中心招募，通过电子问卷和简短的访谈收集的数据包括参与者的社会人口学、健康状况和生活方式、听力和认知功能等方面的信息。研究者还进行了一系列的身体测量，包括血压、动脉超声、视力、身体成分、握力、骨密度以及心电图等；同时采集了参与者的生物样本，包括血液、尿液和唾液。除评估中心访问收集的数据之外，UKB还包含通过网络问卷收集的数据，如在线24 h膳食回顾问卷等。

以文献统计及筛查指南中的变量为基准筛选UKB中的行为生活相关变量,作为本研究的主要观察指标。

参考国内外肺癌筛查指南和高质量肺癌危险因素研究文献[5-27],确定本研究高危人群识别标准。以“肺肿瘤”“肺癌”“发病”“危险”“风险”“模型”“预测”“评价”为中文检索词,以“lung cancer”“lung neoplasms”“lung tumor”“lung carcinoma”“risk”“assessment”“prediction”“score”“model”“paradigm”为英文检索词,分别检索中国知网、万方数据知识服务平台及Pubmed和Web of science数据库的相关文献。时间截止至2023年5月30日,语种限定为中文和英文。最后,统计文献中纳入模型的危险因素。

在不同的资料库中,我们一共找到了2 869篇与之有关的文章,其中中文有357篇,英文有2 512篇。剔除重复、非标准和综述等后,共得到43篇文章。结合国内外肺癌筛查指南,及UKB数据统计资料情况,最终确定本研究分析的肺癌相关危险因素:年龄、性别、体质量指数(body mass index, BMI)、种族、学历、收入、体力活动情况、吸烟状态、饮酒状态、饮酒频率、烹制蔬菜摄入量、沙拉/生食蔬菜摄入量、新鲜水果摄入量、肉类摄入量、奶酪摄入量、焦虑情绪、载脂蛋白A、载脂蛋白B、高密度脂蛋白、低密度脂蛋白、总胆固醇、三酰甘油、癌症家族史、烟草暴露。具体流程见图1。

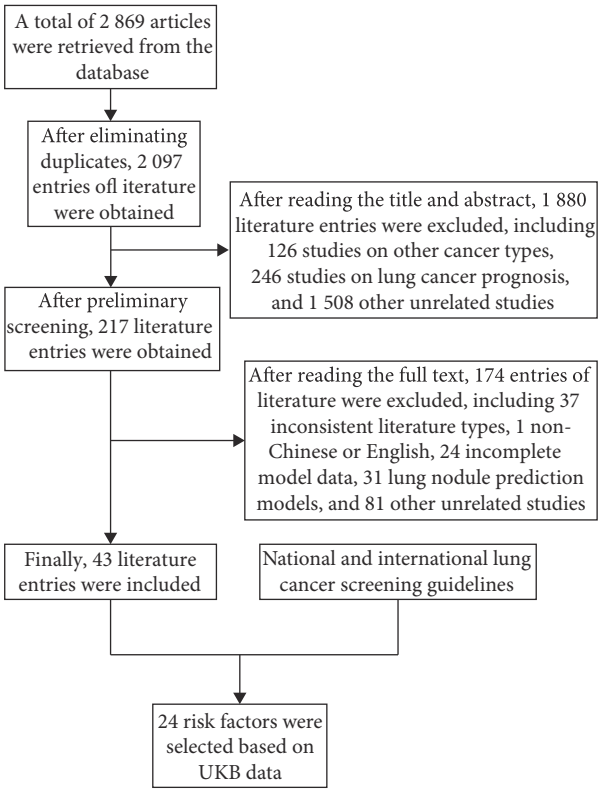

图 1 危险因素筛选流程

Fig 1 Risk factor screening process

1.3 统计学方法

所有统计学分析使用R(4.0.3)进行。定性资料以频数和百分比表示,定量资料以$\bar{x} \pm s$表示,以单因素Cox回归分析组间差异。采用Cox比例风险回归法分析影响肺癌发生的生活行为相关危险因素,自变量筛选用逐步回归法,通过比较赤池信息准则(Akaike information criterion, *AIC*)来选择合适的组合。计算不同指标的风险比(hazard ratio, *HR*)和95%可信区间(confidence interval, *CI*),构建多因素肺癌风险预测模型,使用Schoenfeld残差检验模型中包含的协变量的等比例性,最终选择等比例假设的最优拟合模型。

多因素Cox比例风险回归考虑生存时间,将人群按7∶3的比例随机分为训练集和验证集,使用训练集建立模型,并用验证集对模型性能进行内部验证。使用受试者工作特征(ROC)曲线的曲线下面积(*AUC*)评估模型的效能,$P<0.05$为差异有统计学意义。人群风险评估的方法:根据训练集得出的发病概率,将人群按照发病概率的0% ~ < 25%、25% ~ < 75%、75% ~ 100%分为低风险、中风险及高风险人群,再用验证集人群进行验证评估,即纳入验证人群中总的发病人数,分别计算各风险人群中发病人数占总发病人数的百分比。

2 结果

2.1 研究对象统计学特征及Cox回归分析

研究对象进入队列时未患癌症,453 558人在平均7.79年的随访期间,共发现2 330例肺癌患者,随访的主要结局为肺癌的发生情况。表1总结了参与者的基线特征及单因素Cox回归分析,研究对象的平均年龄为56.52岁,性别男女比例为1.23,结果显示年龄、性别、BMI、学历、收入、种族、体力活动情况、吸烟状态、饮酒状态、饮酒频率、沙拉/生食蔬菜摄入量、新鲜水果摄入量、肉类摄入量、奶酪摄入量、焦虑情绪、载脂蛋白A、高密度脂蛋白、低密度脂蛋白、总胆固醇、三酰甘油、癌症家族史、烟草暴露是肺癌的影响因素($P<0.05$),其余因素无显著性。

将指标纳入逐步回归分析,筛选出10个自变量:年龄、BMI、学历、收入、体力活动情况、吸烟状态、饮酒频率、新鲜水果摄入量、癌症家族史、烟草暴露,进行Cox比例风险回归分析,使用训练集建立模型,并将模型应用于验证集对其判别能力进行了内部验证,结果显示8个自变量(除BMI和新鲜水果摄入量外)均是肺癌的影响因素($P<0.05$)。

2.2 ROC曲线及人群风险评估

训练集样本量为317490人,验证集样本量为136 068

表 1　统计学特征及Cox回归分析

Table 1　Statistical characteristics and Cox regression analysis

Variable	Lung cancer/case, *n*=2 330	No lung cancer/case, *n*=451 228	Incidence/‰	Univariable Cox		Multivariate Cox	
				HR (95% *CI*)	*P*	*HR* (95% *CI*)	*P*
Age/yr.							
40-49	91	111 462	0.82	Reference		Reference	
50-64	1 347	259 088	5.17	6.55 (5.06-8.46)	<0.001	5.43 (3.39-8.69)	<0.001
⩾65	892	80 678	10.94	14.46 (11.14-18.76)	<0.001	11.36 (6.95-18.56)	<0.001
Sex							
Female	1 045	240 686	4.32	Reference			
Male	1 285	210 542	6.07	1.43 (1.3-1.58)	<0.001		
Body mass index/(kg/m^2)							
<18.5	52	4 537	11.33	Reference		Reference	
18.5-24.9	724	145 402	4.95	0.47 (0.33-0.66)	<0.001	3.46 (0.48-24.77)	0.22
25.0-29.9	971	191 225	5.05	0.48 (0.34-0.67)	<0.001	3.11 (0.44-22.23)	0.26
>29.9	583	110 064	5.27	0.48 (0.34-0.68)	<0.001	2.61 (0.36-18.73)	0.34
Qualifications							
College or university degree	362	147 542	2.45	Reference		Reference	
A levels/AS levels or equivalent	181	50 323	3.58	1.62 (1.31-1.99)	<0.001	1.66 (1.23-2.23)	<0.001
O levels/GCSEs or equivalent	424	94 697	4.46	1.87 (1.58-2.21)	<0.001	1.57 (1.22-2.03)	<0.001
CSEs or equivalent	77	24 866	3.09	1.28 (0.95-1.72)	0.105	1.11 (0.65-1.88)	0.71
NVQ, or HND, or HNC, or the equivalent	207	29 622	6.94	2.95 (2.41-3.62)	<0.001	1.71 (1.22-2.38)	<0.001
Other professional qualifications	117	22 960	5.07	2.06 (2.41-2.65)	<0.001	1.55 (1.06-2.25)	0.02
Missing data	962	81 218	11.71				
Income/(£/year)							
<18 000	826	85 352	9.58	Reference		Reference	
18 000-30 999	564	96 737	5.80	0.62 (0.54-0.70)	<0.001	1.02 (0.79-1.32)	0.90
31 000-51 999	299	101 496	2.94	0.32 (0.27-0.37)	<0.001	0.87 (0.66-1.16)	0.35
52 000-100 000	146	80 277	1.82	0.19 (0.16-0.24)	<0.001	0.63 (0.44-0.89)	<0.001
>100 000	39	21 455	1.81	0.17 (0.11-0.26)	<0.001	0.83 (0.49-1.41)	0.50
Missing data	456	65 911	6.87				
Ethnicity							
White	2 250	423 934	5.28	Reference			
Not white	70	25 677	2.72	0.46 (0.34-0.62)	<0.001		
Missing data	10	1 617	6.15				
International Physical Activity Questionnaires activity group							
Low	417	68 539	6.05	Reference		Reference	
Moderate	687	148 113	4.62	0.78 (0.67-0.90)	0.001	0.75 (0.58-0.95)	0.02
High	639	147 235	4.32	0.73 (0.63-0.85)	<0.001	0.61 (0.47-0.79)	<0.001
Missing data	587	87 341	6.68				
Smoking status							
Never	308	249 020	1.24	Reference		Reference	
Previous	1 029	153 457	6.66	5.70 (4.89-6.64)	<0.001	3.74 (2.98-4.68)	<0.001
Current	975	47 071	20.29	16.93 (14.51-19.76)	<0.001	3.23 (1.98-5.29)	<0.001
Missing data	18	1 680	10.60				
Cooked vegetables intake/(tablespoon/d)							
<3	1 177	230 290	5.08	Reference			
3	604	120 535	4.99	0.98 (0.87-1.10)	0.691		
>3	506	94 436	5.33	1.08 (0.96-1.23)	0.195		
Missing data	43	5 967	7.15				
Salad/Raw vegetables intake/(tablespoon/d)							
<3	1 657	305 853	5.39	Reference			
3	270	62 888	4.27	0.80 (0.69-0.93)	0.004		
>3	345	76 357	4.50	0.83 (0.72-0.95)	0.007		
Missing data	58	6 130	9.37				
Fresh fruit intake/(pieces/d)							
<3	1 651	286 932	5.72	Reference		Reference	
3	352	89 765	3.91	0.66 (0.57-0.76)	<0.001	0.80 (0.61-1.03)	0.08
>3	298	72 558	4.09	0.75 (0.65-0.87)	<0.001	1.05 (0.82-1.36)	0.69
Missing data	29	1 973	14.49				
Meat intake/ (tablespoon/d)							
<1	725	178 138	4.05	Reference			
1	677	131 150	5.14	1.21 (1.06-1.37)	0.003		
>1	923	140 841	6.51	1.59 (1.41-1.78)	<0.001		
Missing data	5	1 099	4.53				
Cheese intake/(tablespoon/d)							
<1	509	88 452	5.72	Reference			
1	543	94 305	5.72	1.04 (0.90-1.20)	0.635		
>1	1 198	257 268	4.64	0.85 (0.75-0.96)	0.011		
Missing data	80	11 203	7.09				
Alcohol drinker status							
Never	78	20 245	3.84	Reference			
Previous	186	15 949	11.53	3.26 (2.34-4.54)	<0.001		
Current	2 063	414 475	4.95	1.45 (1.09-1.93)	0.010		
Missing data	3	559	5.34				
Alcohol drinking status/(day/week)							
>4	577	91 413	6.27	Reference		Reference	
1-4	974	221 206	4.38	0.69 (0.61-0.78)	<0.001	0.79 (0.63-1.00)	0.04
<1	776	138 186	5.58	0.88 (0.77-1.00)	0.052	1.06 (0.81-1.38)	0.67
Missing data	3	423	7.04				

续表 1

Variable	Lung cancer/case, *n*=2 330	No lung cancer/case, *n*=451 228	Incidence/‰	Univariable Cox		Multivariate Cox	
				HR (95% *CI*)	*P*	*HR* (95% *CI*)	*P*
Worries/anxious feelings							
No	1 023	191 389	5.32	Reference			
Yes	1 229	247 351	4.94	0.89 (0.81-0.98)	0.021		
Missing data	78	12 488	6.21				
Apolipoprotein A							
Low	9	663	13.39	Reference			
Moderate	1 974	375 222	5.23	0.38 (0.17-0.85)	0.019		
High	42	12 017	3.48	0.26 (0.11-0.62)	0.003		
Missing data	305	63 326	4.79				
Apolipoprotein B							
Low	2	255	7.78	Reference			
Moderate	1 697	325 484	5.19	0.90 (0.13-6.41)	0.918		
High	472	96 954	4.84	0.83 (0.12-5.94)	0.857		
Missing data	159	559	221.45				
High-density lipoprotein							
Low	500	70 339	7.06	Reference			
Moderate	971	184 553	5.23	0.73 (0.65-0.84)	<0.001		
High	562	135 067	4.14	0.60 (0.52-0.69)	<0.001		
Missing data	297	61 269	4.82				
Low-density lipoprotein							
Low	155	15 239	10.07	Reference			
Moderate	710	117 309	6.02	0.65 (0.52-0.80)	<0.001		
High	1 313	291 491	4.48	0.48 (0.39-0.59)	<0.001		
Missing data	152	27 189	5.56				
Total cholesterol							
Low	13	1 181	10.89	Reference			
Moderate	916	145 186	6.27	0.56 (0.29-1.09)	0.088		
High	1 253	278 430	4.48	0.40 (0.21-0.78)	0.007		
Missing data	148	26 431	5.57				
Triacylglycerol							
Low	21	6 735	3.11	Reference			
Moderate	1 123	248 222	4.50	1.37 (0.83-2.24)	0.215		
High	1 036	169 510	6.07	1.75 (1.07-2.88)	0.026		
Missing data	150	26 761	5.57				
Family history of cancer							
No	1 862	392 298	4.72	Reference		Reference	
Yes	452	56 907	7.88	1.75 (1.55-1.97)	<0.001	1.65 (1.30-2.09)	<0.001
Missing data	16	2 023	7.85				
Tobacco exposure							
No	1 033	322 997	3.19	Reference		Reference	
Yes	396	87 133	4.52	1.47 (1.28-1.68)	<0.001	1.33 (1.07-1.65)	0.01
Missing data	901	41 098	21.45				

A levels: Advanced Level qualifications; AS levels: Advanced subsidiary levels; O levels: General Certificate of Education Ordinary Level; GCSEs: General Certificate of Secondary Education; CSEs: Certificate of Secondary Education; NVQ: National Vocational Qualification; HND: higher National Diploma; HNC: Higher National certificate.

人。见图2。针对不同研究时间节点的time-ROC曲线分析结果显示，该模型训练集预测肺癌发生的一年、五年、十年*AUC*分别为0.825(95%*CI*: 0.797 ~ 0.843)、0.785(95%*CI*: 0.772 ~ 0.793)、0.777(95%*CI*: 0.768 ~ 0.789)；验证集预测肺癌发生的一年、五年、十年*AUC*分别为0.857(95%*CI*: 0.846 ~ 0.868)、0.782(95%*CI*: 0.778 ~ 0.786)、0.765(95%*CI*: 0.762 ~ 0.769)。

根据训练集得到发病概率，用验证集发病总人数

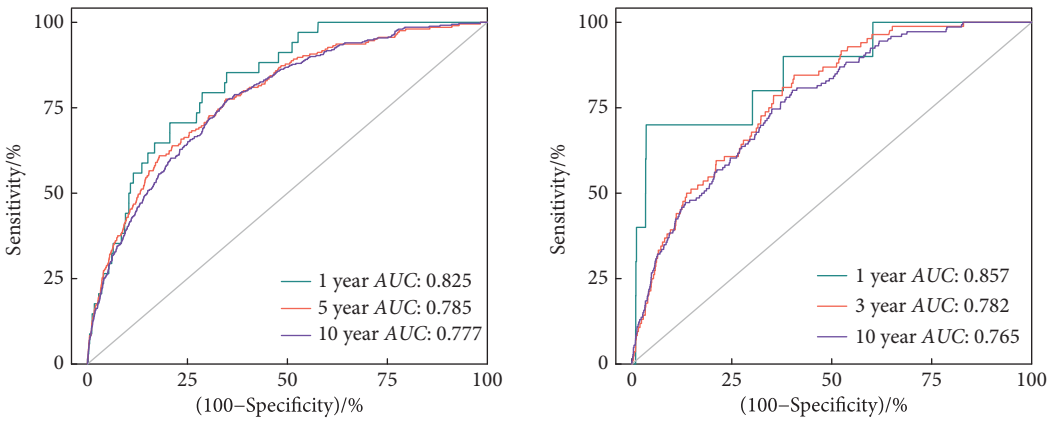

图 2 训练集（左）和验证集（右）ROC曲线分析结果

Fig 2 ROC curve analysis results of the training set (left) and the test set (right)

427进行风险评估。结果显示高风险人群中发病人数达到了总发病人数的68.38%, 即筛查高风险人群即可发现68.38%的未来肺癌发病个体。见表2。

表 2　人群风险评估结果

Table 2　Results of population risk assessment

Risk situation	Number of cases	Proportion of cases
High-risk population	291	68.38%
Moderate-risk population	123	28.58%
Low-risk population	13	3.04%
Total	427	100%

3　讨论

近年来, 国内外学者以不同特征人群为基础, 陆续构建出了多种肺癌风险预测模型。其中, 以 Bach模型[5]、LLP模型[9]等为代表的经典模型表现出了很好的预测效能。一项同样基于UKB构建的肺癌风险预测模型显示, 肺功能指标、肺部疾病史等因素与肺癌风险存在相关性, 而本研究主要聚焦于生活行为相关因素, 提升了干预措施的方便性及可行性[28]。其他研究显示, 吸烟依然是肺癌发病的最主要的危险因素[29], 与此同时, 酒精也会使发生肺癌的风险增大, 结果均与本研究相符, 这提示我们需要采取行之有效的措施, 对人们吸烟和饮酒的行为进行干预。同时有研究表明, 负性生活事件是肺癌发生的重要原因, 其主要原因是长期处于负面情绪中, 通过改变体内的能量代谢进而引起免疫系统的异常, 本研究结果表明焦虑紧张的情绪与肺癌发病风险存在关联性, 因此维持良好的心态, 积极乐观的生活方式, 可以减少肺癌的发生。职业暴露(如烟草、石棉)和污染源环境居住等因素与肺癌发生的风险有很大的关系[30], 本研究证实了烟草暴露与肺癌发生间的关系, 所以需要加强职业环境的监控和管理。

但是, 本研究的局限性如下: 首先, UKB大规模人群队列的参与者来源于一般人群, 肺癌病例数量并不多, 这可能会妨碍我们正确评估预测模型的性能。其次, 如若我们想要使用独立的队列对所建立的模型进行外部验证, 数据集中使用的危险因素定义可能会与此研究中不同, 不易界定的因素(如膳食因素)可能会因为基线调查问卷中的问题措辞不同而有所区别; 膳食因素在地区间的差异同样需要考虑。最后, 由于某些客观原因的限制, 肺癌诊断标准中还应包括CT影像学特征(如肺部结节的类型、直径、数量、毛刺征)、肿瘤标志物、病理学、基因组学等依据[31-33], 需结合其他信息做进一步的研究。此外还需要说明的是, 本研究逐步回归筛选出的变量中, BMI及新鲜水果摄入量在纳入多因素分析时无统计学意义, 可能是由于变量间存在混杂因素或交互效应, 但考虑到其实际意义, 遂在模型中被保留。

综上所述, 本研究结果预期可为将来研究的开展提供参考。目前基于中国人群构建的肺癌预测模型研究类型多为病例对照研究, 且缺乏外部人群的效果评价。在此基础上, 下一步研究可考虑进一步对中国人群进行指标筛选、模型构建与验证, 并对已有的有代表性的肺癌风险预测模型进行有效性对比分析, 也可以考虑将多种因素相结合, 探讨建立具有中国人群特异性的、有针对性的预测模型。同时, 考虑到不同时间和地区的发病差异以及样本规模的有限性, 未来仍需要在更大的人群基础上开展进一步的验证研究, 以评价当前模型是否可在时间和空间上外推; 对模型的不断更新也是必要的, 以构建适用于广泛人群的肺癌风险预测模型, 为不同人群肺癌的早发现、早诊断提供有力的理论依据。该模型具有良好的判别能力, 对临床早发现肺癌, 降低临床肺癌漏诊误诊具有积极意义。

*　　*　　*

作者贡献声明　陈睿琳和王静茹负责论文构思, 陈睿琳和王硕负责数据的编审, 陈睿琳负责正式分析, 索晨负责经费获取、研究项目管理、提供资源和监督指导, 陈睿琳和唐思琦负责调查研究, 陈睿琳负责设计研究方法、可视化和初稿写作, 唐思琦负责软件, 王静茹和王硕负责验证, 陈睿琳、王静茹和索晨负责审读与编辑写作。所有作者已经同意将文章提交给本刊, 且对将要发表的版本进行最终定稿, 并同意对工作的所有方面负责。

利益冲突　所有作者均声明不存在利益冲突

参　考　文　献

[1] SUNG H, FERLAY J, SIEGEL R L, *et al*. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin, 2021, 71(3): 209–249. doi: 10.3322/caac.21660.Epub2021Feb4.

[2] 郑荣寿, 张思维, 孙可欣, 等. 2016年中国恶性肿瘤流行情况分析. 中华肿瘤杂志, 2023, 45(3): 212–220. doi: 10.3760/cma.j.cn112152-20220922-00647.

[3] 赫捷, 陈万青, 李兆, 等. 中国肺癌筛查与早诊早治指南(2022, 北京). 中国肿瘤, 2022, 31(7): 488–527. doi: 10.11735/j.issn.1004-0242.2022.07.A002.

[4] VERNIERI C, NICHETTI F, RAIMONDI A, *et al*. Diet and supplements in cancer prevention and treatment: clinical evidences and future perspectives. Crit Rev Oncol Hematol, 2018, 123: 57–73. doi: 10.1016/j.critrevonc.2018.01.002.

[5] BACH P B, KATTAN M W, THORNQUIST M D, *et al*. Variations in lung cancer risk amongsmokers. J Natl CancerInst, 2003, 95(6): 470–478. doi: 10.1093/jnci/95.6.470.

[6] SPITZ M R, HONG W K, AMOS C I, *et al*. A risk model for prediction of lung cancer. J Natl Cancer Inst, 2007, 99(9): 715–726. doi: 10.1093/jnci/djk153.

[7] SPITZ M R, ETZEL C J, DONG Q, *et al*. An expanded risk prediction model for lung cancer. Cancer Prev Res (Phila), 2008, 1(4): 250–254. doi: 10.1158/1940-6207.capr-08-0060.

[8] El-ZEIN R A, LOPEZ M S, D'AMELIO A M, *et al*. The cytokinesis-blocked micronucleus assay as a strong predictor of lung cancer: extension of a lung cancer risk prediction model. Cancer Epidemiol Biomarkers Prev, 2014, 23(11): 2462–2470. doi: 10.1158/1055-9965.EPI-14-0462.

[9] CASSIDY A, MYLES J P, Van TONGEREN M, *et al*. The LLP risk model: an individual risk prediction model for lung cancer. Br J Cancer, 2008, 98(2): 270–276. doi: 10.1038/sj.bjc.6604158.

[10] RAJI O Y, AGBAJE O F, DUFFY S W, *et al*. Incorporation of a genetic factor into an epidemiologic model for prediction of individual risk of lung cancer: the Liverpool Lung Project. Cancer Prev Res (Phila), 2010, 3(5): 664–669. doi: 10.1158/1940-6207.CAPR-09-0141.

[11] MARCUS M W, CHEN Y, RAJI O Y, *et al*. Llpi: Liverpool lung project risk prediction model for lung cancer incidenc. Cancer Prev Res (Phila), 2015, 8(6): 570–575. doi: 10.1158/1940-6207.capr-14-0438.

[12] MARCUS M W, RAJI O Y, DUFFY S W, *et al*. Incorporating epistasis interaction of genetic susceptibility single nucleotide polymorphisms in a lung cancer risk prediction model. Int J Oncol, 2016, 49(1): 361–370. doi: 10.3892/ijo.2016.3499.

[13] ETZEL C J, KACHROO S, LIU M, *et al*. Development and validation of a lung cancer risk prediction model for African-Americans. Cancer Prev Res (Phila), 2008, 1(4): 255–265. doi: 10.1158/1940-6207.capr-08-0082.

[14] SPITZ M R, AMOS C I, LAND S, *et al*. Role of selected genetic variants in lung cancer risk in African Americans. J Thorac Oncol, 2013, 8(4): 391–397. doi: 10.1097/JTO.0b013e318283da29.

[15] TAMMEMAGI C M, PINSKY P F, CAPORASO N E, *et al*. Lung cancer risk prediction: prostate, lung, colorectal and ovarian cancer screening trial models and validation. J Natl Cancer Inst, 2011, 103(13): 1058–1068. doi: 10.1093/jnci/djr173.

[16] TAMMEMAGI M C, LAM S C, MCWILLIAMS A M, *et al*. Incremental value of pulmonary function and sputum DNA image cytometry in lung cancer risk prediction. Cancer Prev Res (Phila), 2011, 4(4): 552–561. doi: 10.1158/1940-6207.CAPR-10-0183.

[17] TAMMEMAGI M C, KATKI H A, HOCKING W G, *et al*. Selection criteria for lung-cancer screening. N Engl J Med, 2013, 368(8): 728–736. doi: 10.1158/1940-6207.capr-10-0183.

[18] HOGGART C, BRENNAN P, TJONNELAND A, *et al*. A risk model for lung cancer incidence. Cancer Prev Res (Phila), 2012, 5(6): 834–846. doi: 10.1158/1940-6207.CAPR-11-0237.

[19] CHARVAT H, SASAZUKI S, SHIMAZU T, *et al*. Development of a risk prediction model for lung cancer: the Japan public health center-based prospective study. Cancer Sci, 2018, 109(3): 854–862. doi: 10.1111/cas.13509.

[20] YOUNG R P, HOPKINS R J, HAY B A, *et al*. Lung cancer susceptibility model based on age, family history and genetic variants. PLoS One, 2009, 4(4): e5302. doi: 10.1371/journal.pone.0005302.

[21] MAISONNEUVE P, BAGNARDI V, BELLOMI M, *et al*. Lung cancer risk prediction to select smokers for screening CT--a model based on the italian cosmos trial. Cancer Prev Res (Phila), 2011, 4(11): 1778–1789. doi: 10.1158/1940-6207.capr-11-0026.

[22] LI H, YANG L, ZHAO X, *et al*. Prediction of lung cancer risk in a Chinese population using a multifactorial genetic model. BMC Med Genet, 2012, 13: 118. doi: 10.1186/1471-2350-13-118.

[23] PARK S, NAM B H, YANG H R, *et al*. Individualized risk prediction model for lung cancer in Korean men. PLoS One, 2013, 8(2): e54823. doi: 10.1371/journal.pone.0054823.

[24] WANG X, MA K, CUI J, *et al*. An individual risk prediction model for lung cancer based on a study in a Chinese population. Tumori, 2015, 101(1): 16–23. doi: 10.5301/tj.5000205.

[25] 朱猛, 程阳, 戴俊程, 等. 基于全基因组关联研究的中国人群肺癌风险预测模型. 中华流行病学杂志, 2015, 36(10): 1047–1052. doi: 10.3760/cma.j.issn.0254-6450.2015.10.002.

[26] WU X, WEN C P, YE Y, *et al*. Personalized risk assessment in never, light, and heavy smokers in a prospective cohort in Taiwan. Sci Rep, 2016, 6: 36482. doi: 10.1038/srep36482.

[27] WANG X, MA K, CHI L, *et al*. Combining telomerase reverse transcriptase genetic variant rs2736100 with epidemiologic factors in the prediction of lung cancer susceptibility. J Cancer, 2016, 7(7): 846–853. doi: 10.7150/jca.13437.

[28] MULLER D C, JOHANSSON M, BRENNAN P. Lung cancer risk prediction model incorporating lung function: development and validation in the UK Biobank prospective cohort study. J Clin Oncol, 2017, 35(8): 861–869. doi: 10.1200/JCO.2016.69.2467.

[29] WYNDER E L. Tobacco as a cause of lung cancer: some reflections. Am J Epidemiol, 1997, 146(9): 687–694. doi: 10.1093/oxfordjournals.aje.a009342.

[30] OLSSON A C, GUSTAVSSON P, KROMHOUT H, *et al*. Exposure to diesel motor exhaust and lung cancer risk in a pooled analysis from case-control studies in Europe and Canada. Am J Respir Crit Care Med, 2011, 183(7): 941–948. doi: 10.1164/rccm.201006-0940OC.

[31] CHEN W Q, ZHENG R S, BAADE P D, *et al*. Cancer statistics in China, 2015. CA Cancer J Clin, 2016, 66(2): 115–132. doi: 10.3322/caac.21338.

[32] 任冠华, 范亚光, 赵永成, 等. 低剂量螺旋CT肺癌筛查研究进展. 中国肺癌杂志, 2013, 16(10): 553–558.

[33] CRUCITTI P, GALLO I F, SANTORO G, *et al*. Lung cancer screening with low dose CT: experience at Campus Bio-Medico of Rome on 1500 patients. Minerva Chir, 2015, 70(6): 393–399.

(2023 – 06 – 30收稿, 2023 – 08 – 30修回)

编辑 吕 熙