·基因组学与肿瘤风险预测·

基于全基因组关联研究的中国人群肺癌风险预测模型

朱猛 程阳 戴俊程 谢兰 靳光付 马红霞 胡志斌 师咏勇 林东昕 沈洪兵

【摘要】 **目的** 联合使用遗传因素和吸烟信息构建中国汉族人群的肺癌风险预测模型。**方法** 基于中国汉族人群全基因组关联研究(GWAS)数据,根据样本地区来源将样本分为训练集(南京与上海:1 473名病例 *vs.* 1 962名对照)和测试集(北京与武汉:858名病例 *vs.* 1 115名对照)。系统整理已报道肺癌易感位点,在训练集中用逐步后退法筛选具有独立效应的位点,并通过加权法估算个体遗传得分用于建模。在训练集中分别构建基于吸烟信息、遗传得分和联合使用吸烟与遗传信息的3种风险预测模型(吸烟模型、遗传效应模型和联合模型),并根据受试者工作特征(ROC)曲线、曲线下面积(AUC)、净分类指数(NRI)和整体鉴别指数(IDI)评价模型对肺癌风险预测的效能。对于构建的模型,进一步在测试集中进行验证。**结果** 在训练集中,联合模型、吸烟模型和遗传效应模型AUC分别为0.69(0.67~0.71)、0.65(0.63~0.66)和0.60(0.59~0.62)。在训练集和测试集中联合模型的风险预测效能高于吸烟模型或遗传模型,差异有统计学意义($P<0.001$)。重分类结果显示,联合模型与吸烟模型相比,在训练集中NRI增加4.57%(2.23%~6.91%),IDI增加3.11%(2.52%~3.69%)。在测试集中,NRI和IDI分别增加2.77%和3.16%。**结论** 遗传得分可以显著提高肺癌传统风险模型的预测效能。联合使用遗传因素和吸烟信息构建的中国汉族人群肺癌风险预测模型可用于筛选中国汉族人群中肺癌发病的高危人群。

【关键词】 肺癌;全基因组关联研究;风险预测模型

Genome-wide association study based risk prediction model in predicting lung cancer risk in Chinese *Zhu Meng*[1], *Cheng Yang*[1], *Dai Juncheng*[1], *Xie Lan*[2], *Jin Guangfu*[1], *Ma Hongxia*[1], *Hu Zhibin*[1], *Shi Yongyong*[3], *Lin Dongxin*[4], *Shen Hongbing*[1]. 1 *Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing* 211166, China; 2 *Medical Systems Biology Research Center, Tsinghua University School of Medicine*; 3 *Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education, Shanghai Jiao Tong University*; 4 *State Key Laboratory of Molecular Oncology, Cancer Institute and Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College*

Corresponding author: *Shen Hongbing*, Email: hbshen@njmu.edu.cn

This work was supported by grants from the Key Program of National Natural Science Foundation of China (No. 81230067) *and Priority Academic Program for the Development of Jiangsu Higher Education Institutions* (*Public Health and Preventive Medicine*).

【**Abstract**】 **Objective** To evaluate the predictive power of risk model by combining traditional epidemiological factors and genetic factors. **Methods** Our previous GWAS data of lung cancer in Chinese were used in training set (Nanjing and Shanghai: 1 473 cases *vs.* 1 962 control) and testing set (Beijing and Wuhan: 858 cases *vs.* 1 115 control). All the single nucleotide polymorphisms (SNPs) associated with lung cancer risk were systematically selected and stepwise logistic regression analysis was used to select independent factors in the training set. The wGRS (weighted genetic score) was further used to calculate genetic risk score. To evaluate the contribution of the genetic factors, 3 risk models were established by using the training set, i.e. smoking model (based on smoking status), genetic risk model (based on genetic risk score) and combined model (based on smoke and genetic risk score). The predictability of the models were evaluated by the areas under the receiver operating characteristic (ROC) curves, area under curve (AUC), net reclassification improvement (NRI) and integrated discrimination index (IDI). Besides, the results were further

DOI:10.3760/cma.j.issn.0254-6450.2015.10.002

基金项目:国家自然科学基金重点项目(81230067);江苏高校优势学科建设工程专项资金(公共卫生与预防医学)

作者单位:211166 南京医科大学公共卫生学院流行病与卫生统计学系(朱猛、程阳、戴俊程、靳光付、马红霞、胡志斌、沈洪兵);清华大学医学院医学系统生物学研究中心(谢兰);上海交通大学上海市精神卫生中心重点实验室(师咏勇);中国医学科学院北京协和医院肿瘤医院 分子肿瘤学国家重点实验室(林东昕)

通信作者:沈洪兵, Email:hbshen@njmu.edu.cn

verified in the testing set. **Results** In the training set, it was found that the AUC of the smoking, genetic risk and combined models were 0.65 (0.63–0.66), 0.60 (0.59–0.62) and 0.69 (0.67–0.71), respectively. Compared with combined model, the predictive power of other two models significantly declined, the difference was statistically significant ($P<0.001$). Furthermore, compared with the smoking model, the NRI of the combined model increased by 4.57% (2.23%–6.91%) and IDI increased by 3.11% (2.52%–3.69%) in the training set, the difference was statistically significant ($P<0.001$). Similarly, in the testing set NRI increased by 2.77%, the difference was not statistically significant ($P=0.069$), and IDI increased by 3.16%, the difference was statistically significant ($P<0.001$). **Conclusion** This study showed that combining 14 genetic variants with traditional epidemiological factors could improve the predictive power of risk model for lung cancer. The model could be used in the screening of high-risk population of lung cancer in Chinese and provide evidence for the early diagnosis and treatment of lung cancer.

【**Key words**】 Lung cancer; Genome-wide association study; Risk prediction model

2012年中国肺癌新发病例数约为65.3万，标化发病率为36.1/10万，居我国男性恶性肿瘤发病率的首位、女性恶性肿瘤发病率的第二位；由肺癌导致的死亡病例数约为59.7万，占所有恶性肿瘤相关死亡的27.1%，在男性和女性中均居于首位[1]。

依据个人基因信息为癌症和其他疾病制定个体化医疗方案，从而采取个体化的预防策略，是在预防医学领域推动“精准医学”的重要发展方向。自2008年以来，世界范围内开展了多项肺癌相关的全基因组关联研究（GWAS），鉴别出20多个区域的40多个位点与肺癌易感性相关[2-13]。利用GWAS提示的肺癌易感位点信息可能改善基于传统风险因素预测肺癌发病风险的能力。为了验证该假设，本研究基于中国汉族人群GWAS数据库，使用来源于南京与上海汉族人群的样本作为训练集，北京与武汉汉族人群的样本作为测试集，评价遗传信息对肺癌风险预测模型的改善能力，并构建中国汉族人群的肺癌风险预测模型，为中国汉族人群肺癌个体发病预测提供有效的评估工具。

对象与方法

1. 研究对象：基于前期中国汉族人群GWAS数据库[13]，选择2 331名肺癌病例和3 077名健康对照，其中1 473名病例与1 962名对照来源于南京和上海，858名病例与1 115名对照来源于北京和武汉。本研究使用南京与上海作为训练集，用于构建肺癌风险预测模型；使用北京与武汉作为测试集，用于验证肺癌风险预测模型的预测效能。

2. GWAS数据集准备：所有样本均采用美国Affymetrix公司的Affymetrix Genome-Wide Human SNP Array 6.0芯片进行基因分型。分型数据进行严格的质量控制以剔除不合格位点和样本。为了增加数据集覆盖度，对GWAS数据集进行基因型填补（imputation）。首先采用SHAPEIT2软件构建单倍型[14]，然后采用IMPUTE2软件填补未分型位点的基因型[15]，填补以千人基因组计划发布的单倍型数据为参照[16]。

3. 遗传位点选择：系统收集整理GWAS Catalog收录的所有$P<5\times10^{-8}$的肺癌易感位点[17]，剔除在中国人群中最小等位基因频率（MAF）<0.05的位点，对于存在连锁不平衡的位点（$r^2>0.5$），保留P值较小的位点。对于满足上述条件的位点，在训练集样本中采用逐步回归（后退法）筛选，至最终模型中所有位点均有统计学意义（$P<0.05$）。

4. 统计学建模：采用多元logistic回归构建肺癌风险预测模型，根据logistic回归公式，推导出个体发病概率公式：

$$\mathrm{logit}P=\mathrm{logit}\left(\frac{P}{1-P}\right)=\beta_0+\sum_{i=1}^{n}\beta_i x_i$$

$$P=\frac{\exp^{(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}{1+\exp^{(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}=\frac{e^{\mathrm{logit}P}}{1+e^{\mathrm{logit}P}}$$

建模流程见图1。首先在训练集中基于吸烟量构建肺癌传统因素风险预测模型（吸烟模型）；然后基于遗传位点使用加权法（wGRS）估算个体遗传得分，构建肺癌遗传风险预测模型（遗传效应模型）；最后同时结合吸烟和遗传位点信息构建多因素肺癌风险预测模型（联合模型）。加权法是指在对危险因素合并时，考虑因素本身的单独效应，以logistic回归计算得到的变量OR值进行加权得到每个变量的遗传得分，将得分相加后纳入模型，具体公式：

$$\sum_{i=1}^{n}\beta_i X_i$$

式中，n为危险因素的个数，X为危险因素，β为权重，即经logistic计算后得到的每个危险因素的OR值。既往研究表明，加权法的预测效能优于其他方法[18]。

根据受试者工作特征（ROC）曲线下面积（AUC）、净分类指数（NRI）和整体鉴别指数（IDI）评

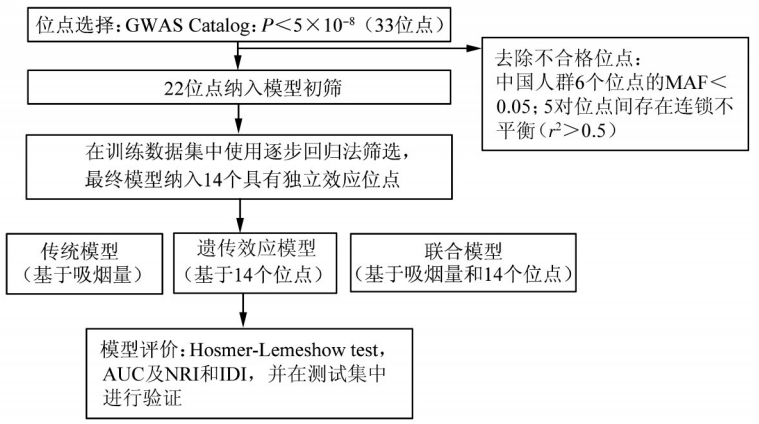

图1 肺癌发病风险预测模型建立流程

价上述3种模型的预测效能，并进一步在测试集中进行验证。所有分析利用R3.1.3软件完成。

结　果

1. 样本基线特征：研究样本的基线特征见表1。在训练样本与测试样本中，吸烟状态与吸烟量差异均有统计学意义（$P<0.001$）。

2. 单核苷酸多态性（SNP）位点选择：截至2015年4月20日，GWAS Catalog共收录21项肺癌全基因组易感性关联研究，共包含41个独立SNP位点，其中33位点$P<5\times10^{-8}$，在中国人群中有6个位点MAF<0.05，5对位点间存在高度连锁不平衡（$r^2>0.5$），最终有22个肺癌相关位点进入模型初筛。在训练样本中，经逐步回归分析后，最终14个位点进入模型（表2）。在模型纳入的14个位点中，除了rs938682（15q25.1，CHRNA3–CHRNA5）发现于欧美人群之外，其余13个位点均来源于亚洲人群。

3. 不同模型预测肺癌发生风险：吸烟模型显示，在训练集中，与不吸烟者相比，轻度吸烟者患肺癌风险增加1.21倍（调整$OR=2.21$，95%CI：1.76～2.78），重度吸烟者患肺癌的风险增加3.76倍（调整$OR=4.76$，95%CI：3.88～5.84）。遗传效应模型显示，根据14个位点计算个体的遗传得分，并根据对照的得分将样本分为4组，随着得分增加，个体患肺癌风险显著增加（$P=2.04\times10^{-23}$）。联合模型显示，根据对照危险得分将样本分为4组，组间风险增加趋势变得更加明显（$P=1.13\times10^{-54}$），得分最高组患肺癌的风险是得分最低组的5.73倍（95%CI：4.52～7.26）。这一变化趋势在测试集中得到了验证（表3）。

4. 模型预测能力与效度检验：联合模型、吸烟模型和遗传效应模型AUC分别为0.69、0.65和0.60，见表4。与吸烟模型和遗传效应模型相比，联合模型的预测效能较高（$P<0.001$），见图2。Hosmer-Lemeshow检验表明3种模型的拟合度均较好（$P>0.05$）。这种趋势在测试集中得到了验证（图2），吸烟模型和遗传效应模型的AUC均为0.61（95%CI：0.58～0.63），联合模型的AUC增加为0.65（95%CI：0.62～0.67），预测效能提高（$P<0.001$）。但是基于训练样本建立的联合模型和吸烟模型，在测试样本中拟合度较差（$P<0.05$）。进一步对年龄、性别进行分层分析，发现不同层间该趋势保持一致（图3）。在中国大部分女性为不吸烟者的基本国情下，吸烟模型预测效能差，AUC为0.54（95%CI：0.53～0.56），而联合模型的预测效能为0.64（95%CI：0.61～0.67）。

5. 阳性界值与重分类：根据联合模型，使用Youden法在训练样本中确定的最佳阳性界值为0.4，

表1 研究样本的基线特征

特　征	训练样本		测试样本		合　计	
	病例(n=1 473)	对照(n=1 962)	病例(n=858)	对照(n=1 115)	病例(n=2 331)	对照(n=3 077)
年龄(岁，$\bar{x}\pm s$)	60.08±10.30	59.35±9.74	60.00±10.23	62.45±9.18	60.05±9.93	60.47±9.93
≤60	712(48.34)	1 081(55.10)	430(50.12)	440(39.46)	1 142(48.99)	1 521(49.43)
>60	761(51.66)	881(44.90)	428(49.88)	675(60.54)	1 189(51.01)	1 556(50.57)
性别						
男	1 057(71.76)	1 214(61.88)	654(76.22)	872(78.21)	1 711(73.40)	2 086(67.80)
女	416(28.24)	748(38.12)	204(23.78)	243(21.79)	620(26.60)	991(32.20)
吸烟状态						
吸烟	741(50.31)	636(32.42)	511(59.56)	447(40.09)	1 252(53.71)	1 083(35.20)
戒烟	168(11.41)	83(4.23)	86(10.02)	143(12.83)	254(10.90)	226(7.34)
不吸烟	564(38.29)	1 243(63.35)	261(30.42)	525(47.09)	825(35.39)	1 768(57.46)
吸烟量(包/年，$\bar{x}\pm s$)	41.43±26.86	30.96±20.28	44.81±29.64	32.35±19.88	42.77±28.04	31.59±20.11
≤25	254(27.94)	327(45.48)	166(27.81)	232(39.32)	420(27.89)	559(42.70)
>25	655(72.06)	392(54.52)	431(72.19)	358(60.68)	1 086(72.11)	750(57.30)

注：括号外数据为人数，括号内数据为构成比（%）

表2　逐步回归模型纳入14个肺癌相关SNP位点的基本信息及在训练集中的关联

位点	染色体位置(碱基改变)	染色体区域	基因	MAF 病例	MAF 对照	*OR*值(95%*CI*)	*P*值[a]	研究报道
rs2131877	chr3:194858374(G→A)	3q29	*C3orf21*	0.42	0.44	0.87(0.78～0.97)	1.51×10^{-2}	Yoon等[8]
rs4488809	chr3:189356261(T→C)	3q28	*TP63*	0.52	0.47	1.27(1.14～1.41)	1.96×10^{-5}	Hu等[13]；Lan等[3]
rs2853677	chr5:1287194(G→A)	5p15	*TERT-CLPTM1L*	0.41	0.38	1.18(1.05 –1.33)	4.52×10^{-3}	Shiraishi等[7]
rs2895680	chr5:146644115(C→T)	5q32	*PPP2R2B-STK32A-DPYSL3*	0.33	0.28	1.23(1.09～1.38)	4.84×10^{-4}	Dong等[2]
rs465498	chr5:1325803(A→G)	5p15	*TERT-CLPTM1L*	0.12	0.15	0.77(0.66～0.91)	1.42×10^{-3}	Hu等[13]
rs9387478	chr6:117786180(C→A)	6q22.2	*ROS1-DCBLD1*	0.48	0.50	0.89(0.80～0.99)	3.55×10^{-2}	Lan等[3]
rs1663689	chr10:9025195(T→C)	10p14	*GATA3*	0.39	0.42	0.88(0.79～0.98)	1.68×10^{-2}	Dong等[2]
rs7086803	chr10:114498476(G→A)	10q25.2	*VTI1A*	0.31	0.27	1.21(1.07～1.36)	2.33×10^{-2}	Lan等[3]
rs12296850	chr12:100820085(A→G)	12q23.1	*SLC17A8*	0.22	0.25	0.79(0.70～0.90)	3.94×10^{-4}	Dong等[4]
rs753955	chr13:24293859(A→G)	13q12	*MIPEP-TNFRSF19*	0.32	0.29	1.24(1.10～1.40)	4.00×10^{-4}	Hu等[13]
rs938682	chr15:78896547(G→A)	15q25.1	*CHRNA3-CHRNA5*	0.41	0.43	0.89(0.80～0.99)	3.59×10^{-2}	Broderick等[6]
rs7216064	chr17:65898809(G→A)	17q24.3	*BPTF*	0.37	0.40	0.90(0.81～1.01)	6.72×10^{-2}	Shiraishi等[7]；Lan 等[3]
rs4809957	chr20:52771171(A→G)	20q13.2	*CYP24A1*	0.38	0.35	1.16(1.04～1.30)	8.80×10^{-3}	Dong等[4]
rs17728461	chr22:30598552(C→G)	22q12	*MTMR3-HORMAD2*	0.20	0.17	1.24(1.08～1.43)	2.75×10^{-3}	Hu等[13]

注：[a]调整年龄、性别、吸烟量(年包)及第一主成分

表3　风险预测模型的基本情况

模型	病例(%)	对照(%)	调整*OR*值(95%*CI*)[c]	*P*值[c]	趋势性*P*值[c]
训练样本					
吸烟(包/年)					
不吸烟	564(38.29)	1 243(63.35)	1.00		
≤25	254(17.24)	327(16.67)	2.21(1.76～2.78)	1.17×10^{-11}	
>25	655(44.46)	392(19.98)	4.76(3.88～5.84)	1.17×10^{-50}	3.39×10^{-51}
遗传效应[a]					
0(<Q_{25})	208(14.12)	492(25.08)	1.00		
1(Q_{25}～)	313(21.25)	493(25.13)	1.45(1.16～1.82)	1.17×10^{-3}	
2(Q_{50}～)	396(26.88)	487(24.82)	2.00(1.61～2.49)	5.18×10^{-10}	
3(>Q_{75})	556(37.75)	490(24.97)	2.75(2.23～3.40)	7.20×10^{-21}	2.04×10^{-23}
联合[b]					
0(<Q_{25})	156(10.59)	493(25.12)	1.00		
1(Q_{25}～)	214(14.53)	490(24.97)	1.40(1.10～1.79)	6.12×10^{-3}	
2(Q_{50}～)	338(22.95)	489(24.92)	2.35(1.86～2.96)	6.45×10^{-13}	
3(>Q_{75})	765(51.93)	490(24.97)	5.73(4.52～7.26)	2.05×10^{-47}	1.13×10^{-54}
测试样本					
吸烟(包/年)					
不吸烟	261(30.42)	525(47.09)	1.00		
≤25	166(19.35)	232(20.81)	2.06(1.54～2.76)	9.66×10^{-7}	
>25	431(50.23)	358(32.11)	3.88(2.99～5.03)	1.67×10^{-24}	4.47×10^{-25}
遗传效应[a]					
0(<Q_{25})	155(18.07)	281(25.20)	1.00		
1(Q_{25}～)	145(16.90)	279(25.02)	0.94(0.70～1.26)	6.83×10^{-1}	
2(Q_{50}～)	215(25.06)	278(24.93)	1.34(1.01～1.77)	4.29×10^{-2}	
3(>Q_{75})	343(39.98)	277(24.84)	2.35(1.80～3.06)	3.67×10^{-10}	1.43×10^{-12}
联合[b]					
0(<Q_{25})	87(10.14)	279(25.02)	1.00		
1(Q_{25}～)	189(22.03)	280(25.11)	2.27(1.66～3.11)	2.61×10^{-7}	
2(Q_{50}～)	209(24.36)	277(24.84)	3.37(2.42～4.68)	4.58×10^{-13}	
3(>Q_{75})	373(43.47)	279(25.02)	6.60(4.76～9.15)	1.18×10^{-29}	5.82×10^{-31}

注：[a]根据对照中遗传得分四分类；[b]根据对照中危险得分四分类；[c]对于吸烟和联合模型，调整年龄、性别；对于遗传效应，调整年龄、性别、吸烟量

此时预测的灵敏度和特异度分别为64.70%和64.93%。以0.4为阳性界值，进一步评价吸烟模型和联合模型对病例和对照重分类的影响(表5)，结果显示，在训练集中联合模型NRI增加4.57%(95%*CI*：2.23%～6.91%)，IDI增加3.11%(95% *CI*：2.52%～3.69%)，差异均有统计学意义($P<0.001$)。在测试样本中，NRI增加2.77%，但差异无统计学意义($P=0.069$)，而IDI增加了3.16%，差异有统计学意义($P<0.001$)。

讨　论

肺癌发病率和死亡率均位于我国恶性肿瘤的首位，且起病隐匿、恶性程度高、病程进展快、易复发和转移、预后较差。因此，通过对肺癌高危人群进行筛查，达到早诊早治是降低肺癌发病率和死亡率的理想手段。近年来，研究者们通过GWAS这一最新的研究手段，发现了一系列与肺癌易感性密切相关的遗传位点。本研究主要探讨了如何有效的利用这些研究成果，改善肺癌传统的风险预测模型，进一步鉴别肺癌高危人群，以减少疾病筛查的经济和健康负担。

目前，国内外已发表多种肺癌发病风险评估模型，如Bach模型、LLP

表4　模型校准度及AUC比较

模型	AUC	95%*CI*	P_{AUC}值[a]	模型校准度	
				χ^2值	*P*值[b]
训练集					
联合	0.69	0.67～0.71	1.000	5.17	0.739
吸烟	0.65	0.63～0.66	＜0.001	7.07	0.530
遗传效应	0.60	0.59～0.62	＜0.001	1.51	0.993
测试集					
联合	0.65	0.62～0.67	1.000	29.72	＜0.001
吸烟	0.61	0.58～0.63	＜0.001	22.84	0.004
遗传效应	0.61	0.58～0.63	＜0.001	8.62	0.376

注：[a] AUC比较，以联合模型为参照，单纯考虑吸烟和遗传因素预测效能均显著降低；[b]基于Hosmer-Lemeshow 拟合优度检验

模型和Etzel模型等[19-21]。纳入的危险因素分别包括吸烟、戒烟、石棉接触史、痰细胞学检查及呼吸性疾病史等。Bach等[19]通过使用年龄、吸烟（吸烟量、吸烟年数和戒烟年数）和石棉暴露等危险因素在18 172名吸烟或戒烟人群中构建肺癌风险预测模型，AUC可达到0.7～0.9，然而该模型缺乏外部数据的验证，且目前石棉已经淘汰出日常生活，该模型已经失去其原有的使用价值。而LLP和Etzel模型[20-21]主要使用年龄、吸烟、肿瘤家族史和职业暴露等因素，AUC仅在0.55～0.70间。Hoggart等[22]使用初始吸烟年龄、吸烟量和10个职业环境暴露因素，根据初始吸烟年龄分层构建风险预测模型，预测

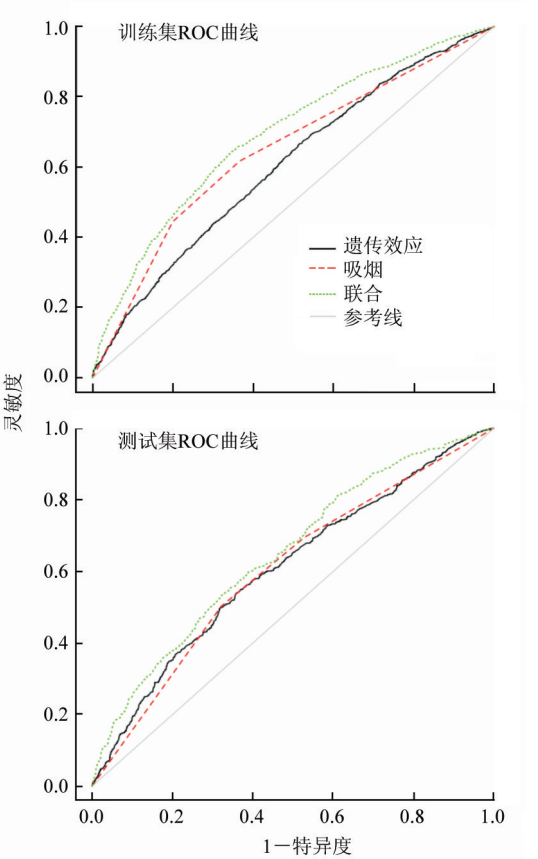

图2　基于GWAS结果构建的中国人群肺癌风险预测模型预测能力

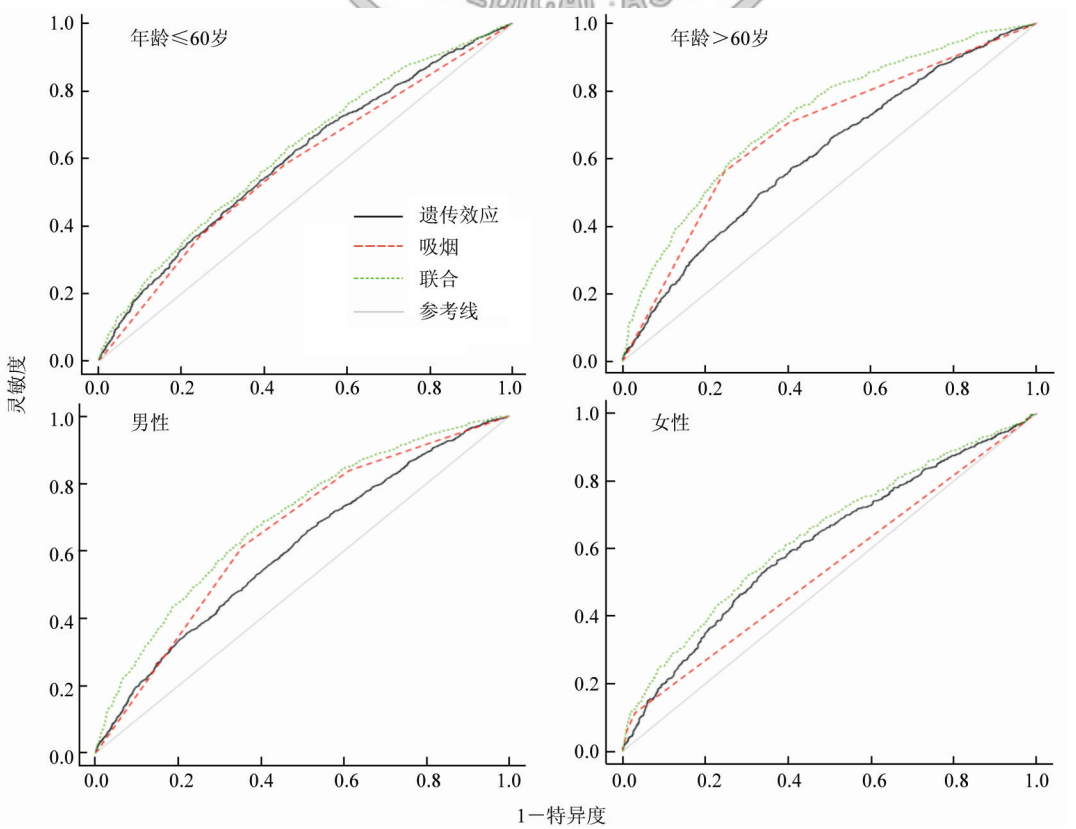

图3　中国人群肺癌风险预测模型分层分析的ROC曲线

表5　联合模型与吸烟模型重分类比较

传统模型	联合模型[a]		
	0～0.4	0.4～1.0	重分类比例 (%)
训练样本（对照）			
0～0.4	1 139	104	8
0.4～1.0	135	584	19
训练样本（病例）			
0～0.4	454	110	20
0.4～1.0	66	843	7
训练样本（合并）			
0～0.4	1 593	214	12
0.4～1.0	201	1 427	12
NRI			4.57(2.23～6.91), $P<0.001$
IDI			3.11(2.52～3.69), $P<0.001$
测试样本（对照）			
0～0.4	472	53	10
0.4～1.0	80	510	14
测试样本（病例）			
0～0.4	213	48	18
0.4～1.0	45	552	8
测试样本（合并）			
0～0.4	685	101	13
0.4～1.0	125	1 062	11
NRI			2.77(−0.22～5.76), $P=0.069$
IDI			3.16(2.36～3.95), $P<0.001$

注：[a] 根据Youden法确定最佳阳性界值0.4进行重分类

效能较好，AUC为0.84（0.81～0.88），该研究尝试在模型构建时引入GWAS发现的2个遗传位点（5p15、15q25），然而仅有小部分样本拥有分型信息，引入遗传信息并没有显著改善模型的预测效能。对于上述模型，多只适用于吸烟或戒烟者，并不适用于不吸烟者，且上述模型主要基于欧美人群构建，能否直接应用于中国汉族人群，尚有待进一步验证。

传统模型主要基于流行病调查数据（如年龄、吸烟状态和职业暴露等），在信息收集的过程中难免存在信息偏倚，与此相比，基于遗传易感位点的遗传得分可以稳定存在，检测方法稳定且可重复，是用于风险预测的可靠资源。本研究基于前期中国汉族人群GWAS数据库，以南方样本作为训练集用于构建肺癌风险预测模型、北方样本作为测试集验证肺癌风险预测模型的预测效能，系统评价了遗传信息对于传统肺癌风险预测模型的改善能力。经过系统筛选，最终选择14个既往GWAS报道的SNPs用于构建遗传得分，联合模型在训练集和测试集中均可以显著改善单纯基于吸烟信息的模型。在分层分析中，不同组间这种趋势均得以验证，提示模型稳定可靠。

参　考　文　献

[1] Bray F, Ren JS, Masuyer E, et al. Global estimates of cancer prevalence for 27 sites in the adult population in 2008[J]. Int J Cancer, 2013, 132(5): 1133–1145.

[2] Dong J, Hu ZB, Wu C, et al. Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population[J]. Nat Genet, 2012, 44(8): 895–899.

[3] Lan Q, Hsiung CA, Matsuo K, et al. Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia[J]. Nat Genet, 2012, 44(12): 1330–1335.

[4] Dong J, Jin GF, Wu C, et al. Genome-wide association study identifies a novel susceptibility locus at 12q23.1 for lung squamous cell carcinoma in han chinese[J]. PLoS Genet, 2013, 9(1): e1003190.

[5] Hu ZB, Xia YK, Guo XJ, et al. A genome-wide association study in Chinese men identifies three risk loci for non-obstructive azoospermia[J]. Nat Genet, 2012, 44(2): 183–186.

[6] Broderick P, Wang YF, Vijayakrishnan J, et al. Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study[J]. Cancer Res, 2009, 69(16): 6633–6641.

[7] Shiraishi K, Kunitoh H, Daigo Y, et al. A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population[J]. Nat Genet, 2012, 44(8): 900–903.

[8] Yoon KA, Park JH, Han J, et al. A genome-wide association study reveals susceptibility variants for non-small cell lung cancer in the Korean population[J]. Hum Mol Genet, 2010, 19(24): 4948–4954.

[9] Miki D, Kubo M, Takahashi A, et al. Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations[J]. Nat Genet, 2010, 42(10): 893–896.

[10] Wang YP, Lu Y, Zhang Y, et al. The draft genome of the grass carp (Ctenopharyngodon idellus) provides insights into its evolution and vegetarian adaptation[J]. Nat Genet, 2015, 47(6): 625–631.

[11] Landi MT, Chatterjee N, Yu K, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma[J]. Am J Hum Genet, 2009, 85(5): 679–691.

[12] Deng QF, Guo H, Dai JC, et al. Imputation-based association analyses identify new lung cancer susceptibility variants in CDK6 and SH3RF1 and their interactions with smoking in Chinese populations[J]. Carcinogenesis, 2013, 34(9): 2010–2016.

[13] Hu ZB, Wu C, Shi YY, et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese[J]. Nat Genet, 2011, 43(8): 792–796.

[14] Delaneau O, Marchini J. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel[J]. Nat Commun, 2014, 5: 3934.

[15] Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies[J]. PLoS Genet, 2009, 5(6): e1000529.

[16] 1 000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1 092 human genomes[J]. Nature, 2012, 491(7422): 56–65.

[17] Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations[J]. Nucleic Acids Res, 2014, 42(Database issue): D1001–1006.

[18] Dai J, Hu Z, Jiang Y, et al. Breast cancer risk assessment with five independent genetic variants and two risk factors in Chinese women[J]. Breast Cancer Res, 2012, 14(1): R17.

[19] Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers[J]. J Natl Cancer Inst, 2003, 95(6): 470–478.

[20] Cassidy A, Myles JP, van Tongeren M, et al. The LLP risk model: an individual risk prediction model for lung cancer[J]. Br J Cancer, 2008, 98(2): 270–276.

[21] Spitz MR, Hong WK, Amos CI, et al. A risk model for prediction of lung cancer[J]. J Natl Cancer Inst, 2007, 99(9): 715–726.

[22] Hoggart C, Brennan P, Tjonneland A, et al. A risk model for lung cancer incidence[J]. Cancer Prev Res (Phila), 2012, 5(6): 834–846.

（收稿日期：2015–06–15）
（本文编辑：万玉立）