论 著

# 前列腺癌3种早期诊断预测模型效能比较

李　翔[1]，李姝婷[1]，陈　容[1]，骆仕俐[1]，张　溪[1]，李远盛[1]，许望东[2]，张俊辉[1]
西南医科大学：1.公共卫生学院；2.科技处（泸州 646000）

【摘　要】**目的** 采用logistic回归、决策树和Lagrangian支持向量机（Lagrangian Sopport Vector Machine，LSVM）三种方法构建前列腺癌的早期诊断预测模型，并比较三种模型的预测效能，为前列腺癌的早期诊断提供理论支持。**方法** 数据来源于国家临床医学科学数据中心（301医院）的《前列腺肿瘤预警数据集》，将清洗整理后的数据按7∶3的比例随机分成训练集和测试集，基于训练集数据采用单因素Logistic回归筛选前列腺癌的关联因素，并建立多因素Logistic回归分析、LSVM和随机森林模型三个前列腺癌早期诊断预测模型，用测试集数据验证三个模型的预测准确性并用ROC曲线对三种模型进行评价比较。**结果** 单因素Logistic分析筛选出13项具有统计学意义的指标，包括年龄、肌酸激酶同工酶、甘油三酯、磷脂、游离PSA、总PSA、钙、血清尿酸、载脂蛋白A1、载脂蛋白B、载脂蛋白C2、载脂蛋白C3、载脂蛋白E。多因素Logistic分析筛选出4个有统计学意义的变量：年龄、肌酸激酶同工酶、游离PSA、总PSA。LSVM模型筛选出10项预测因子，按重要性由高至低分别是：总PSA、年龄、载脂蛋白A1、磷脂、载脂蛋白B、甘油三酯、血清尿酸、游离PSA、肌酸磷酸同工酶、载脂蛋白E。随机森林模型筛选出10项预测因子，按重要程度排序由高至低依次为：载脂蛋白C3、磷脂、游离PSA、载脂蛋白B、载脂蛋白E、钙、血清尿酸、载脂蛋白A1、载脂蛋白C2、肌酸激酶同工酶。多因素非条件Logistic回归、LSVM模型和随机森林模型分析的AUC分别为0.895（0.876，0.913）、0.918（0.902，0.934）、0.724（0.688，0.760）。**结论** LSVM模型预测效果最好，多因素Logistic回归模型预测效果尚可，随机森林的预测效果不佳。

【关键词】前列腺癌；风险预测；Logistic回归；LSVM模型；随机森林模型；ROC曲线
【中图分类号】R737.25　　文献标志码　A　　DOI：10.3969/j.issn.2096-3351.2023.04.011

# A comparative study of three risk prediction model of prostate cancer

*LI Xiang[1], LI Shuting[1], CHEN Rong[1], LUO Shili[1], ZHANG Xi[1], LI Yuansheng[1], XU Wangdong[2], ZHANG Junhui[1]*
1. *School of Public Health, Southwest Medical University, Luzhou 646000, China*; 2. *Department of Science and Technology, Southwest Medical University, Luzhou 646000, China*

【**Abstract**】**Objective** Logistic regression, decision tree, and LSVM were used to construct a predictive model for early diagnosis of prostate cancer, and the predictive performance of the three models was compared to provide theoretical support for the early diagnosis of prostate cancer. **Methods** The data were obtained from the prostate tumor data set of the national clinical medical science data center (301 Hospital). The cleaned data were randomly divided into a training set and testing set according to the proportion of 7∶3. Based on the training set data, single factor unconditional Logistic regression was applied to screen the related factors of prostate cancer, and unconditional multivariate Logistic regression analysis and Lagrangian Support Vector Machine (LSVM) model and random forest model were established risk prediction models for prostate cancer. The testing data were used to verify the accuracy of the three models, and ROC curves were used to evaluate their performances. **Results** The results of single factor logistic analysis showed that 13 indicators, including age, creatine kinase isoenzyme, triglyceride, phospholipid, free PSA, total PSA, calcium, serum uric acid, apolipoprotein A1, apolipoprotein B, apolipoprotein C2, apolipoprotein C3 and apolipoprotein E, were statistically significant. Four variables were selected by multivariate Logistic analysis: age, creatine kinase isoenzyme, free-PSA, total-PSA. LSVM model screened out 10 predictive factors, according to the importance from high to low: total-PSA, age, apolipoprotein A1, phospholipid, apolipoprotein B, triglyceride, serum uric acid, free-PSA, creatine phosphate isoenzyme, and apolipoprotein E. The random forest model selected ten predictive factors, with the order of importance being apolipoprotein C3, phospholipid, free-PSA, apolipoprotein B, apolipoprotein E, calcium, serum uric acid, apolipoprotein A1, apolipoprotein C2, and creatine kinase isoenzyme. The area under the curve (AUC) of multivariate unconditional logistic regression, LSVM model, and random forest model were 0.895(0.876, 0.913), 0.918(0.902, 0.934) and 0.742(0.688, 0.760), respectively. **Conclusion** LSVM model had the best predictive performance, followed by the multivariate logistic regression model, and the random forest model had the poorest predictive performance.

【**Key words**】Prostatic carcinoma; Risk prediction; Logistic regression; LSVM model; Random forest model; ROC curve

**基金项目**：全国统计科学研究项目（2021LZ31）；泸州市政府-西南医科大学公共卫生学院创新团队项目（SPH18001）
**通信作者**：张俊辉，E-mail：vzjh960500@163.com
**引用本文**：李翔，李姝婷，陈容，等.前列腺癌3种早期诊断预测模型效能比较[J].西南医科大学学报.2023，46(4)：330-335.DOI：10.3969/j.issn.2096-3351.2023.04.011

前列腺癌（prostate cancer）是指发生在前列腺的上皮性恶性肿瘤，是最常见的男性泌尿生殖系统恶性肿瘤之一，占男性恶性肿瘤病死率的第2位[1-2]。据《全球癌症统计报告》数据显示[3]2020年中国前列腺癌标化发病率为10.2 / 10万，标化死亡率为4.6 / 10万。目前，中国前列腺癌的总体发病率和死亡率低于世界水平，但前列腺癌的发病率和死亡率仍在不断升高[4]，癌症顺位前移，其疾病负担有向年轻人群转移的趋势[5]。前列腺癌患者在发病早期往往缺乏典型的临床表现。早期诊断指标主要是前列腺特异性抗原（prostate special antigen，PSA），该指标具有器官特异性而非肿瘤特异性，在前列腺增生、前列腺炎和其他前列腺良性疾病中都可能会升高[6]，导致前列腺癌过度诊断和过度治疗的风险很高[7]。因此，本研究首先采用单因素Logistic回归初步筛选了前列腺癌的预测因子，并据此建立了Logistic回归、Lagrangian支持向量机（Lagrangian Support Vector Machine, LSVM）和随机森林模型三种前列腺癌早期诊断预测模型，再采用ROC曲线对其预测效能进行比较，以探讨三种模型在前列腺癌早期诊断中的应用价值，为前列腺癌早期诊断提供理论支持。

## 1 资料与方法

### 1.1 数据来源

数据来源于临床医学科学数据中心（301医院）的《前列腺肿瘤预警数据集》[8]。数据集包含基本信息表、诊断表、检查信息表、病理信息表、PSA信息表、生化检查表、导尿信息表、血常规信息表、放疗信息表、手术情况表、性腺激素表、药物信息表、膀胱镜信息表等。包括患者编号、检查结果标签、年龄、身高、体重、骨钙素、载脂蛋白A2、快速微量尿白蛋白/肌酐测定、磷脂、血清血白蛋白、α1球蛋白、α2球蛋白、β1球蛋白、β2球蛋白、γ球蛋白、游离PSA、总PSA、钠、脑利钠肽前体、尿钠、钙、尿钙等共46项信息。

### 1.2 数据清洗

将该数据中前列腺癌预测的结果变量重新定义为有无前列腺癌（1 = 有，0 = 无）。把前列腺增生、前列腺癌、同时有前列腺增生和前列腺癌三类重新赋值为两类，其中前列腺增生归为无前列腺癌（0 = 无），前列腺癌和同时有前列腺癌和前列腺增生归为有前列腺癌（1 = 有）。由于原始数据中含有文字的数据不能导入SPSS 25.0，会导致数据的缺失，因此将数据中含有的文字删除，再对数据进行核实补充。将每个变量由小到大排序，根据逻辑关系判断各变量的极大值和极小值是否为异常值，结果发现有个别异常值存在。由于存在少量数据缺失和个别异常值，快速微量尿白蛋白/肌酐测定、骨钙素、脑利钠肽前体、α1球蛋白、α2球蛋白、β1球蛋白、β2球蛋白、γ球蛋白、尿酸、尿钠、尿钙、尿氯化物、尿磷、尿肌酐、肌钙蛋白T等19项变量被去除，剩余25项变量。对信息清洗整理后分析得到数据库共有2 987例个案，其中前列腺癌个案222例，占7.4%；无前列腺癌个案2 765例，占92.6%。考虑到样本含量较大，数据有少量缺失对模型比较结果影响有限，因此本文未作数据填补。

### 1.3 纳入和排除标准

本文研究对象的纳入标准为：①各项指标齐全；②诊断信息明确。排除标准为：①重复个案；②变量缺失过多。根据上诉纳入排除标准删除个案324例，最终确定2 663例研究对象，其中患前列腺癌个案数220例，对照组为未患前列腺癌个案数2 443例。

### 1.4 变量定义

变量赋值参考《前列腺癌筛查专家共识》[9]和《中国成人血脂异常防治指南》[10]。

### 1.5 统计学分析

将整理好的2 663例研究对象按7:3的比例随机划分为训练集和测试集，其中用来训练模型的训练集一共1 864例，用来测试模型的测试集一共799例。

采用SPSS 25.0及SPSS Modeler client进行数据整理及分析，基于训练集数据采用单因素Logistic回归筛选前列腺癌的关联因素，并据此建立多因素Logistic回归分析、LSVM模型和随机森林模型三个前列腺癌早期诊断预测模型，用验证集数据验证三个模型的预测准确性并用ROC曲线下面积（area under curve，AUC）对三种模型的预测性能进行比较[11]。

1.5.1 LSVM模型介绍　支持向量机（SVM）是一类有监督学习的广义线性分类器，是由VAPNIK等提出的一种机器学习算法，能较好地解决小样本、高维数、非线性和局部极小点等实际问题[12]。SVM包括Proximal支持向量机（简称PSVM）和LSVM等，LSVM的精度优于其他模型，其分类效果在临床实践中也得到了广泛验证。

1.5.2 随机森林模型　随机森林算法是一种包含多个决策树的集成学习方法，以决策树为基学习器，运用Bagging的方法进行集成，能显著提高基学习器的性能，具有算法精度高、能处理大规模数据和减少过拟合等优点，在特征选择和高维数据分析等领域已得到许多关注[13]。

## 2 结果

### 2.1 单因素Logistic回归分析

将训练集中的变量逐个纳入单因素Logistic回归模型，结果显示血清白蛋白、乳酸脱氢酶、碱性磷酸酶、钠、肌酸激酶等无统计学意义（$P > 0.05$）。年龄、磷脂、游离PSA、总PSA、钙、血清尿酸、载脂蛋白A1、载脂蛋白B、载脂蛋白C3等指标有统计学意义（$P < 0.05$），可以作为预测模型的备选指标进一步分析，见表1。

表1　前列腺癌关联因素的单因素非条件Logistic回归分析结果

Table 1　Results of univariate unconditional logistic regression analysis for prostate cancer

| 变量 | 分类 | 频数 | B | S.E. | Wald | *P* | *OR* | *OR*的95%*CI* |
|---|---|---|---|---|---|---|---|---|
| 年龄(岁) | 45～60[a] | 1 342 | | | 172.674 | < 0.001 | | |
| | < 45 | 433 | − 1.055 | 0.531 | 3.952 | 0.047 | 0.348 | 0.123～0.985 |
| | > 60 | 888 | 2.258 | 0.190 | 140.501 | < 0.001 | 9.560 | 6.582～13.886 |
| BMI | 正常[a] | 2 006 | | | 0.000 | 1.000 | | |
| | 偏低 | 110 | 18.772 | 23 205.556 | 0.000 | 0.999 | 142 040 604.709 | 0.000 |
| | 偏高 | 547 | 0.000 | 46 410.911 | 0.000 | 1.000 | 1.000 | 0.000 |
| 载脂蛋白A | 正常[a] | 2 648 | 18.771 | 15 191.569 | 0.000 | 0.999 | 141 957 465.748 | 0.000 |
| 载脂蛋白C2 | 正常[a] | 2 006 | | | 6.847 | 0.033 | | |
| | 偏低 | 110 | − 0.048 | 0.406 | 0.014 | 0.907 | 0.953 | 0.430～2.114 |
| | 偏高 | 547 | − 0.677 | 0.259 | 6.837 | 0.009 | 0.508 | 0.306～0.844 |
| 载脂蛋白E | 正常[a] | 2 239 | | | 7.223 | 0.027 | | |
| | 偏低 | 98 | − 0.994 | 0.724 | 1.882 | 0.170 | 0.370 | 0.090～1.531 |
| | 偏高 | 326 | 0.497 | 0.223 | 4.976 | 0.026 | 1.644 | 1.062～2.543 |
| 磷脂 | < 2.17[a] | 649 | | | 12.387 | 0.006 | | |
| | 2.17～2.46 | 671 | 0.212 | 0.228 | 0.869 | 0.351 | 1.236 | 0.791～1.932 |
| | 2.46～2.79 | 670 | 0.064 | 0.235 | 0.074 | 0.786 | 1.066 | 0.673～1.688 |
| | ⩾ 2.79 | 673 | − 0.731 | 0.286 | 6.540 | 0.011 | 0.481 | 0.275～0.843 |
| 血清白蛋白 | 正常[a] | 52 | − 0.490 | 0.731 | 0.450 | 0.502 | 0.612 | 0.146～2.566 |
| 肌酸激酶同工酶 | 正常[a] | 258 | 0.880 | 0.220 | 15.969 | < 0.001 | 2.411 | 1.566～3.714 |
| 游离PSA | < 0.18[a] | 602 | | | 104.197 | < 0.001 | | |
| | 0.18～0.27 | 680 | − 0.561 | 0.167 | 11.333 | 0.001 | 0.570 | 0.411～0.791 |
| | 0.27～0.43 | 703 | − 2.370 | 0.298 | 63.303 | < 0.001 | 0.093 | 0.052～0.168 |
| | ⩾ 0.43 | 678 | − 1.692 | 0.224 | 56.954 | < 0.001 | 0.184 | 0.119～0.286 |
| 总PSA | 正常[a] | 319 | 2.763 | 0.188 | 215.938 | < 0.001 | 15.844 | 100.961～22.904 |
| 钙 | 正常[a] | 2 143 | − 0.363 | 0.077 | 22.262 | < 0.001 | 0.696 | 0.598～0.809 |
| 无机磷 | 正常[a] | 2 298 | | | 1.454 | 0.483 | | |
| | 偏低 | 332 | 0.237 | 0.199 | 1.426 | 0.232 | 1.268 | 0.859～1.872 |
| | 偏高 | 33 | 0.139 | 0.610 | 0.052 | 0.820 | 1.149 | 0.347～3.801 |
| 游离钙 | 正常[a] | 204 | 0.431 | 0.230 | 3.527 | 0.060 | 1.539 | 0.981～2.415 |
| 乳酸脱氢酶 | 正常[a] | 23 | 0.860 | 0.555 | 2.402 | 0.121 | 2.363 | 0.797～7.007 |
| 血清尿酸 | 正常[a] | 622 | − 0.834 | 0.211 | 15.635 | < 0.001 | 0.435 | 0.287～0.657 |
| 甘油三酯 | 正常[a] | 1 594 | | | 20.770 | < 0.001 | | |
| | 偏低 | 20 | − 0.779 | 1.029 | 0.573 | 0.449 | 0.459 | 0.061～3.450 |
| | 偏高 | 1 049 | − 0.729 | 0.161 | 20.443 | < 0.001 | 0.482 | 0.352～0.662 |
| 高密度脂蛋白胆固醇 | 正常[a] | 866 | | | 10.517 | 0.005 | | |
| | 偏低 | 1 570 | − 0.314 | 0.153 | 4.227 | 0.040 | 0.730 | 0.541～0.985 |
| | 偏高 | 227 | 0.350 | 0.231 | 2.308 | 0.129 | 1.419 | 0.903～2.230 |
| 载脂蛋白A1 | 正常[a] | 917 | | | 100.943 | < 0.001 | | |
| | 偏低 | 1 633 | − 0.847 | 0.234 | 13.106 | < 0.001 | 0.429 | 0.271～0.678 |
| | 偏高 | 113 | − 2.134 | 0.247 | 74.718 | < 0.001 | 0.118 | 0.073～0.192 |
| 载脂蛋白B | 正常[a] | 701 | | | 28.790 | < 0.001 | | |
| | 偏低 | 1 766 | − 0.268 | 0.163 | 2.712 | 0.100 | 0.765 | 0.556～1.052 |
| | 偏高 | 196 | 0.841 | 0.227 | 13.678 | < 0.001 | 2.319 | 1.485～3.621 |
| 载脂蛋白C3 | < 7.34[a] | 716 | | | 8.187 | 0.042 | | |
| | 0.18～0.27 | 683 | 0.185 | 0.212 | 0.763 | 0.383 | 1.203 | 0.794～1.823 |
| | 0.27～0.43 | 644 | 0.498 | 0.203 | 6.039 | 0.014 | 1.645 | 1.106～2.448 |
| | ⩾ 0.43 | 620 | 0.469 | 0.205 | 5.206 | 0.023 | 1.598 | 1.068～2.389 |
| 碱性磷酸酶 | 正常[a] | 2 437 | | | 0.139 | 0.933 | | |
| | 偏低 | 201 | 0.095 | 0.258 | 0.137 | 0.711 | 1.100 | 0.664～1.824 |
| | 偏高 | 25 | − 0.028 | 0.741 | 0.001 | 0.970 | 0.973 | 0.228～4.155 |
| 钠 | 正常[a] | 2 493 | | | 0.053 | 0.974 | | |
| | 偏低 | 21 | 0.156 | 0.747 | 0.044 | 0.835 | 1.169 | 0.270～5.052 |
| | 偏高 | 149 | − 0.028 | 0.310 | 0.008 | 0.928 | 0.972 | 0.530～1.784 |
| 肌酸激酶 | 正常[a] | 2 178 | − 0.063 | 0.179 | 0.124 | 0.725 | 0.939 | 0.660～1.334 |
| 肌酐 | 正常[a] | 2 541 | | | 1.805 | 0.406 | | |
| | 偏低 | 1 | −18.775 | 40 192.970 | 0.000 | 1.000 | 0.000 | 0.000 |
| | 偏高 | 121 | 0.394 | 0.293 | 1.805 | 0.179 | 1.483 | 0.835～2.636 |
| 低密度脂蛋白胆固醇 | 正常[a] | 1 562 | | | 1.644 | 0.439 | | |
| | 偏低 | 439 | − 0.209 | 0.204 | 1.050 | 0.306 | 0.811 | 0.544～1.210 |
| | 偏高 | 661 | − 0.169 | 0.172 | 0.967 | 0.326 | 0.844 | 0.603～1.183 |
| 常数 | | | − 37.814 | 25 796.208 | 0.000 | 0.999 | 0.000 | |

注：a指该变量的对照组。

## 2.2 三种前列腺癌预测模型的建立

**2.2.1 多因素非条件Logistic回归分析** 有序多分类的似然比结果见表2，可见载脂蛋白C3的有序多分类变量似然比结果差异无统计学意义（$P>0.05$），应以分组线性变量形式纳入模型，其余变量以哑变量形式纳入模型，自变量赋值表见表1。将训练集中单因素分析后有统计学意义的变量纳入多因素非条件Logistic回归经逐步向后回归（纳入标准为$P<0.05$，排除标准为$P>0.1$）筛除变量后，筛选出的主要影响因素为：年龄、肌酸激酶同工酶、游离PSA、总PSA，各影响因素$P$值及$OR$值见表3。

**表2 有序多分类变量纳入模型方式检验结果**

Table 2 Test results of inclusion of ordered multi-classification variables into the model

| 变量 | −2lnL | | G | 自由度 | $P$ |
|---|---|---|---|---|---|
| | 哑变量 | 分组线性 | | | |
| 载脂蛋白E | 1 510.606 | 1 505.900 | 4.706 | 1 | < 0.05 |
| 载脂蛋白A1 | 1 506.625 | 1 412.835 | 93.790 | 1 | < 0.05 |
| 载脂蛋白B | 1 514.394 | 1 449.929 | 64.465 | 1 | < 0.05 |
| 载脂蛋白C3 | 1 511.317 | 1 510.080 | 1.237 | 1 | > 0.05 |

**表3 前列腺癌关联因素的多因素非条件Logistic回归分析结果**

Table 3 Results of multivariate unconditional logistic regression analysis for prostate cancer

| 变量 | | B | S.E | Wald | $P$ | $OR$ | $OR$的95%$CI$ |
|---|---|---|---|---|---|---|---|
| 年龄（岁） | 45 ~ 60[a] | | | 80.079 | < 0.001 | | |
| | < 45 | − 0.866 | 0.538 | 2.589 | 0.108 | 0.421 | 0.147 ~ 1.208 |
| | ⩾ 60 | 1.704 | 0.212 | 64.598 | < 0.001 | 5.499 | 3.629 ~ 8.332 |
| 肌酸激酶同工酶 | 正常[a] | 0.554 | 0.217 | 6.518 | 0.011 | 1.74 | 1.137 ~ 2.662 |
| 游离PSA | < 0.18[a] | | | 68.968 | < 0.001 | | |
| | 0.18 ~ 0.27 | − 1.933 | 0.337 | 32.902 | < 0.001 | 0.145 | 0.075 ~ 0.280 |
| | 0.27 ~ 0.43 | − 1.47 | 0.278 | 28.05 | < 0.001 | 0.23 | 0.133 ~ 0.396 |
| | ⩾ 0.43 | − 2.337 | 0.33 | 50.181 | < 0.001 | 0.097 | 0.051 ~ 0.184 |
| 总PSA | 正常[a] | 3.307 | 0.298 | 122.849 | < 0.001 | 27.31 | 15.217 ~ 49.013 |
| 常量 | | − 2.976 | 0.203 | 215.108 | < 0.001 | 0.051 | |

注：a指该变量的对照组。

**2.2.2 LSVM模型** 利用训练集中单因素Logistic回归筛选出的有统计学意义的变量建立LSVM模型。结果显示，总PSA、年龄、载脂蛋白A1、磷脂、载脂蛋白B、甘油三酯、血清尿酸、游离PSA、肌酸激酶同工酶、载脂蛋白E为主要的影响因素。变量重要性排序见图1。

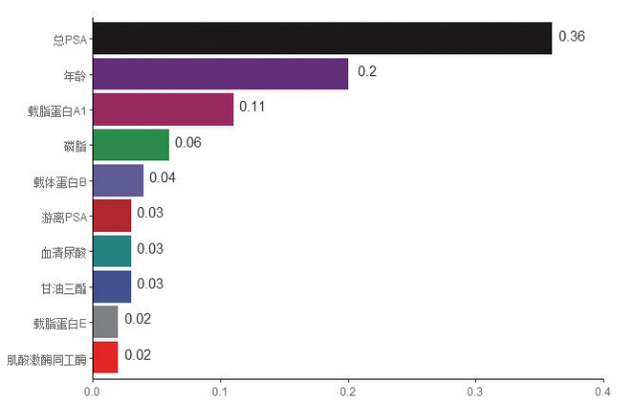



**图1 LSVM模型预测变量重要性排序**

Figure 1 Importance diagram of LSVM predictive variables

**2.2.3 随机森林模型** 利用训练集中单因素Logistic回归筛选出的有统计学意义的变量建立随机森林模型。结果显示，以变量重要性为顺序包括：载脂蛋白C3、磷脂、游离PSA、载脂蛋白B、载脂蛋白E、钙、血清尿酸、载脂蛋白A1、载脂蛋白C2、肌酸激酶同工酶。变量重要性排序如图2。

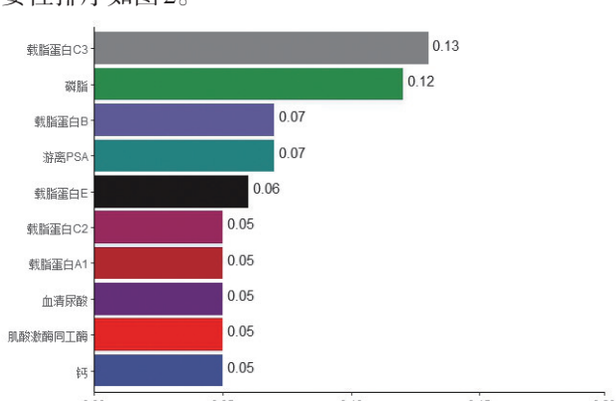



**图2 随机森林模型预测变量重要性排序**

Figure 2 Importance of variables predicted by random forest plot

## 2.3 三种模型的预测效果比较

将建立好的三种模型用于测试集中做预测，结果显示，多因素非条件Logistic回归、LSVM和随机森林的ROC曲线下面积（AUC）分别为：0.895（0.876，0.913）、0.918（0.902，0.934）、0.724（0.688，0.760），详见表4。LSVM模型预测效果最好（AUC > 0.9），而Logistic回归模型和随机森林模型的AUC在0.7 ~ 0.9之间，拟合效果虽不如LSVM但效果仍可接受。预测结果的ROC曲线见图3。

表4 三种模型的ROC曲线下面积(AUC)比较
Table 4 Comparison of area under ROC curve (AUC) among the three models

| 模型 | AUC | S.E. | *P* | AUC的95% *CI* |
|---|---|---|---|---|
| Logistic | 0.895 | 0.009 | < 0.001 | 0.876 ~ 0.913 |
| LSVM | 0.918 | 0.008 | < 0.001 | 0.902 ~ 0.934 |
| 随机森林 | 0.724 | 0.018 | < 0.001 | 0.688 ~ 0.760 |

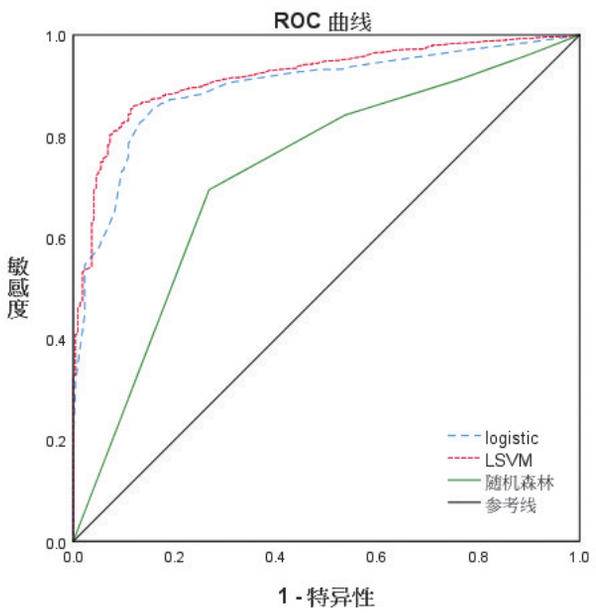



图3 三种模型的ROC曲线比较
Figure 3 Comparison of ROC curves of the three models

## 3 讨论

本研究采用单因素Logistic回归初步筛选了前列腺癌的预测因子，并据此建立了Logistic回归、LSVM和随机森林三种前列腺癌早期诊断预测模型。其中，LSVM模型的AUC最高，提示如果用LSVM来建立前列腺癌早期诊断预测模型准确性可能会更高。三种模型筛选出的变量不全相同，随机森林预测效果欠佳，因此可将Logistic回归模型和LSVM两种方法的结果相互印证补充，结合专业知识确定前列腺癌的重要预测因子。Logistic回归模型和LSVM模型共同筛选出的预测因子为总PSA、游离PSA、年龄、肌酸激酶同工酶，这四个变量可作为前列腺癌的预测因子也在相关研究中得到过验证[14]。

PSA和年龄是公认的前列腺癌预测模型的重要预测因子，我们的研究也显示，PSA和年龄可能是预测模型最重要的两个预测因子。1994年以来，PSA广泛应用于前列腺癌的早期诊断、筛查、预后随访等方面，也是美国食品和药物管理局（FDA）批准的第一个肿瘤生物标志物[15-17]。年龄增大是前列腺癌最重要的不可控因素，随着年龄的增加，患前列腺癌的可能性也随之增加，年龄升高也常常伴随前列腺癌病理Gleason评分以及临床分期的明显升高[18-19]。因此老年人应作为前列腺癌的重点关注人群，在诊断时提高警惕。

本研究结果提示肌酸激酶同工酶CK-MB是前列腺癌的重要预测因子。丁慧等[20]研究发现，前列腺癌患者CK和CK-MB含量较正常组均显著升高，其原因可能是由于人体免疫系统对于肿瘤组织和细胞的攻击从而导致其破裂释放出一定CK及CK-MB含量，因此一部分恶性肿瘤患者会出现血清CK及CK-MB含量升高的体征。

本研究也发现了一些具有争议性的预测因子，如载脂蛋白E、载脂蛋白C2、载脂蛋白C3等。在既往的研究中，LIU等[21]认为E4基因及其等位基因与前列腺癌的发病和预后无关，但FARUK等[22]的一项研究表明，载脂蛋白E影响前列腺癌发生和高Gleason评分的出现，其可能是一个区分前列腺癌的生物标志物，我们的研究也支持了这一观点。

目前国外已有较多基于临床数据建立的前列腺癌早期诊断预测模型，如目前最常使用的两种风险预测模型为：前列腺癌预防试验风险计算器（PCPT-RC）模型、欧洲前列腺癌风险计算器筛查随机研究（ERSPC-RC）模型；此外TOMLINS等[23]基于Logistic回归模型，利用血清PSA、PSAD以及PCA3等预测因子建立了前列腺评分系统。国内关于前列腺癌早期诊断预测模型的研究较少，且这些模型并未纳入如磷脂、载体蛋白等较易获得的生化检查指标；而目前已有研究[24]表明这些生化检查指标可能能够提高前列腺癌早期诊断的特异度和灵敏度。因此本研究在常规前列腺癌诊断预测因子的基础上，加入了这些容易获得的生化检测指标作为备选预测因子。研究结果显示，纳入了生化检查指标的LSVM模型与LEE等[25]建立的SVM前列腺癌诊断模型相比，拥有更高的AUC值。

本研究的不足之处：首先，由于国家临床医学科学数据中心（301医院）提供的《前列腺癌数据集》有部分数据缺失，可能会对结果有一定的影响。另外，我们建立的前列腺癌预测模型只利用了测试集进行内部数据验证，是否可以外推至其他数据集还有待进一步验证。

## 4 结论

本研究利用国家临床医学科学数据中心（301医院）的《前列腺肿瘤数据集》，筛选出年龄、PSA、肌酸激酶同工酶等前列腺癌早期诊断预测因子，并在此基础上构建了基于PSA联合指标的多因素Logistic回归模型、随机森林模型和LSVM前列腺癌早期诊断预测模型。研究结果显示，LSVM模型预测效果最好，多因素Logistic回归模型预测效果尚可，随机森林模型的预测效果不佳。

## 5 参考文献

[1] 赫捷, 陈万青, 李霓, 等. 中国前列腺癌筛查与早诊早治指南(2022, 北京) [J]. 中华肿瘤, 2022, 44(1): 29-53.
[2] BRAY F, FERLAY J, S0ERJOMATARAM I, *et al*. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries [J]. A Cancer Journal for Clinicians, 2018, 68(6): 394-42.
[3] 刘宗超, 李哲轩, 张阳, 等. 2020全球癌症统计报告解读 [J]. 肿瘤综合治疗电子杂志, 2021, 7(2) : 1-13, I0001.
[4] ZHAI Z, ZHENG Y, LI N, *et al*. Incidence and disease burden of prostate cancer from 1990 to 2017: results from the Global Burden of Disease Study 2017[J]. Cancer, 2020, 126(9): 1969-1978.
[5] 付振涛, 郭晓雷, 张思维,等. 2015年中国前列腺癌发病与死亡分析 [J]. 中华肿瘤, 2020, 42(9): 718-722.
[6] 冉俊武, 王伟, 蓝中凯, 等. 前列腺组织学炎症分级对前列腺良性疾病及前列腺癌患者血清PSA的影响 [J]. 当代医学, 2022, 28(5): 98-102.
[7] HENDRIKS R J, VAN OORT I M, SCHALKEN J A. Blood-based and urinary prostate cancer biomarkers: a review and comparison of novel biomarkers for detection and treatment decisions [J]. Prostate Cancer and Prostatic Diseases, 2017, 20(1): 12-19.
[8] 中国人民解放军总医院.前列腺肿瘤预警数据集.国家人口健康科学数据中心数据仓储 PHDA, 2022. https://doi. org/10.12213/11.A0005.201905.000531.
[9] 中国抗癌协会泌尿男生殖系统肿瘤专业委员会前列腺癌学组. 前列腺癌筛查专家共识 [J]. 中华外科杂志, 2017, 55(5): 435-440.
[10] 中国成人血脂异常防治指南修订联合委员会. 中国成人血脂异常防治指南(2016年修订版) [J]. 中华全科医师杂志, 2017, 16(1): 15-35.
[11] 车宏鑫, 王桐, 王伟. 前列腺癌预测模型对比研究 [J]. 数据分析与知识发现, 2021, 5(9): 107-114.
[12] 丁捷, 张大成, 王博文, 等 基于LIBS与化学计量学的植物叶片分类研究 [J]. 光谱学与光谱分析, 2021, 41(2): 606-611.
[13] 汪力纯, 刘水生. 基于混合采样和特征选择的改进随机森林算法研究 [J]. 南京邮电大学学报(自然科学版), 2022, 42(1): 81-89.
[14] 韦雅芹, 程锐璇, 史维俊, 等. 基于数据挖掘的前列腺癌风险预测研究 [J]. 广西医科大学学报, 2020, 37(10): 1847-1851.
[15] 邓姣, 马金鹏, 杨德林. 前列腺癌早期诊断的研究进展 [J]. 现代肿瘤医学, 2022, 30(5): 903-907.
[16] MALODE S J, SHETTI N P. New generation biomarkers for the detection of prostate cancer [J]. Biosensors and Bioelectronics: X, 2022, 12: 100250.
[17] CARLSSON S V, MURATA K, DANILA D C, *et al*. PSA: role in screening and monitoring patients with prostate cancer [M]. Cancer Biomarkers. Amsterdam: Elsevier, 2022: 131-172.
[18] 宋刚, 纪光杰, 张雷, 等. 前列腺癌患者年龄与病理分级关系的研究(附2929例报告) [J]. 中华泌尿外科杂志, 2017, 38(2): 106-109..
[19] RAWLA P. Epidemiology of prostate cancer [J]. World Journal of Oncology, 2019, 10(2): 63-89.
[20] 丁慧, 董振南, 温新宇, 等. 七种恶性肿瘤患者血清CK和CK-MB含量分析 [J]. 标记免疫分析与临床, 2013, (2): 87-89.
[21] LIU H, SHUI I, PLATZ E, *et al*. No association of ApoE genotype with risk of prostate cancer: a nested case control study [J]. Cancer Epidemiology and Prevention Biomarkers, 2015, 24(10): 1632-4.
[22] FARUK Y, GULEC Y S, ASIF Y, *et al*. Apolipoprotein E genotypes in patients with prostate cancer [J]. Anticancer Research, 2016, 36(2): 707-11.
[23] TOMLINS S A, DAY J R, LONIGRO R J, *et al*. Urine TMPRSS2: ERG plus PCA3 for individualized prostate cancer risk assessment [J]. European Urology, 2016, 70(1): 45-53.
[24] GUO H X, JIA X J, LIU H. Based on biomedical index data: Risk prediction model for prostate cancer [J]. Medicine, 2021, 100(17).
[25] LEE H J, HWANG S I, HAN S M, *et al*. Image-based clinical decision support for transrectal ultrasound in the diagnosis of prostate cancer: comparison of multiple logistic regression, artificial neural network, and support vector machine [J]. Eur Radiol, 2010 Jun;20(6):1476-84.

**(利益冲突:无)**

(收稿日期:2022-08-21;修回日期:2022-12-03)

(上接第325页)

[25] CUI J, GONG C Q, CAO B R, *et al*. MicroRNA-27a participates in the pathological process of depression in rats by regulating VEGFA[J]. Experimental and Therapeutic Medicine:2018,15(5): 4349-4355.
[26] WALKE PRACHI B, BANSODE SNEHA B, MORE NIKITA P, *et al*. Molecular investigation of glycated insulin-induced insulin resistance via insulin signaling and AGE-RAGE axis[J]. BBA - Molecular Basis of Disease, 2021,1867(2): 166029.
[27] FERNANDO H K D H, LEUNG C, ANGUS P, *et al*. FRI-295-Advanced glycation end products (AGEs) exacerbate NAFLD progression to liver fibrosis via receptor for AGEs (RAGE) in high fat fed mice[J]. Journal of Hepatology, 70(1): e525.
[28] CHEN S, OKAHARA F, OSAKI N, *et al*. Increased GIP signaling induces adipose inflammation via a HIF-1α-dependent pathway and impairs insulin sensitivity in mice[J]. American Journal of Physiology-Endocrinology and Metabolism, 2015,308(5): E414-E425.
[29] XIE Z P, HUANG S X, XIE S K, *et al*. 2021. Potential correlation between depression-like behavior and the mitogen-activated protein kinase pathway in the rat hippocampus following spinal cord injury[J]. World Neurosurgery, 2021,154: e29-e38.
[30] SU W J, ZHANG Y, CHEN Y, *et al*. NLRP3 gene knockout blocks NF-κB and MAPK signaling pathway in CUMS-induced depression mouse model[J]. Behavioural Brain Research, 2017, 322: 1-8.
[31] 李媛媛, 杨钦河, 张金文, 等.疏肝健脾方药对非酒精性脂肪性肝病大鼠肝细胞TLR4/p38MAPK信号通路的影响[J]. 中国老年学杂志, 2015,35(23): 6649-6653.
[32] HEMMATI F, GHASEMI R, IBRAHIM N M, *et al*. Crosstalk between insulin and toll-like receptor signaling pathways in the central nervous system[J]. Molecular Neurobiology, 2014,50(3): 797-810.
[33] KWON E Y, CHOI M S. Luteolin targets the toll-like receptor signaling pathway in prevention of hepatic and adipocyte fibrosis and insulin resistance in diet-induced obese mice[J]. Nutrients, 2018,10(10): 1415.

**(利益冲突:无)**

(收稿日期:2022-09-12;修回日期:2023-01-08)