

基于深度神经网络的老年肺癌发病风险评估研究^{*}

陈松景 吴思竹

(中国医学科学院/北京协和医学院医学信息研究所 北京 100020)

〔摘要〕 目的/意义 利用深度神经网络开展老年肺癌发病风险评估研究, 为降低肺癌发病率和有效防控肺癌的发生提供理论依据。方法/过程 整合多种相关危险因素, 对整合后的数据进行预处理, 构建基于深度神经网络的肺癌发病风险评估模型, 并运用对比分析法进行模型验证与评价。结果/结论 男女高风险组的肺癌发病率均明显高于低风险组, 男性高风险组肺癌发病率是女性高风险组肺癌发病率的2.02倍, 在男性高风险组进行干预控制对降低肺癌发病率更加显著。本研究不仅有利于老年肺癌的早诊早治, 也为提高老年人生活质量提供了基础支撑。

〔关键词〕 深度神经网络; 老年肺癌; 风险评估

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.12.004

Study on Risk Assessment of Lung Cancer in the Elderly Based on Deep Neural Networks

CHEN Songjing, WU Sizhu

Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

〔Abstract〕 **Purpose/Significance** To Utilize deep neural networks to carry out the risk assessment study of lung cancer in the elderly, and to provide the theoretical basis for reducing the incidence of lung cancer and its effective preventive control. **Method/Process** A variety of related risk factors are integrated, and the integrated data are preprocessed. The deep neural network-based risk assessment model for lung cancer is constructed. And the comparative analysis is used to carry out the model validation evaluation. **Result/Conclusion** The incidence of lung cancer of the high-risk group is significantly higher than the low-risk group. The incidence of lung cancer of the male high-risk group is 2.02 times higher than the female high-risk group. The intervention control of the male high-risk group is more significant for reducing lung cancer incidence. The study is not only beneficial to the early diagnosis and treatment of lung cancer in the elderly, but also provides basic support for improving the quality of life of the elderly.

〔Keywords〕 deep neural networks; elderly lung cancer; risk assessment

1 引言

肺癌在全球范围内发病率高, 且在我国癌症发病率中居首位, 每年新发病例大多集中于老年群

体, 这给积极应对人口老龄化和实现“健康中国2030”目标带来了巨大挑战。因此, 开展肺癌发病风险评估研究, 对降低老年肺癌发病率和防控肺癌的发生具有重要意义。

肺癌的发病涉及复杂多样的危险因素, 如年龄、

〔修回日期〕 2024-10-22

〔作者简介〕 陈松景, 副研究员, 发表论文30余篇; 通信作者: 吴思竹, 研究员。

〔基金项目〕 国家自然科学基金项目(项目编号: 62106286); 中国医学科学院医学与健康科技创新工程项目(项目编号: 2021-I2M-1-057)。

日常生活习惯、吸烟情况、慢性病史等，这些因素中隐藏了大量与肺癌风险评估相关的信息。美国国家癌症研究所主导的国家癌症筛查试验^[1]和荷兰-比利时肺癌筛查试验^[2]通过临床随机对照试验开展肺癌评估筛查研究。瑞典卡罗林斯卡医学院 Nemlander E 等^[3]通过横截面电子调查问卷数据，针对从未吸烟、曾经吸烟和目前吸烟等不同吸烟状况参与者，采用基于随机梯度下降的机器学习方法开展肺癌发病评估预测研究。目前国内外在肺癌发病风险评估方面的研究大多聚焦于横截面数据^[4-5]，或借助人工经验^[6-7]、传统方法^[8-9]等，在高性能算法、纵向数据分析等方面探索相对有限，难以满足老年肺癌发病风险准确评估的需求。本研究采用深度神经网络方法，基于大规模随访人群数据开展老年肺癌发病风险评估及其验证研究，旨在为肺癌的预防控制提供基础依据，并为提升老年人生活质量提供有力支撑。

2 资料与方法

2.1 资料来源

资料来源于美国健康与退休研究数据库 (<https://hrs.isr.umich.edu>)，包括美国 50 岁以上约 20 000 人的代表性数据。基线调查始于 1992 年，随后每两年进行一次追踪。本研究整合 1992—2020 年的随访数据。所涉及的信息主要包括出生年月、性别、身体质量指数、是否患有高血压、癌症病史、肺及气管相关疾病情况、心脏相关疾病情况、是否患有糖尿病、目前吸烟情况、曾经吸烟情况、精神状况、去世年月等。

2.2 方法

基于深度神经网络方法，利用大规模人群队列数

据开展老年肺癌发病风险评估研究。首先，整合多种危险因素的相关数据；其次，对整合后的数据进行预处理；再次，构建基于深度神经网络的肺癌发病风险评估模型；最后，通过统计分析对该模型进行验证，实现老年肺癌发病风险的准确评估。具体研究过程涵盖数据整合、数据预处理、肺癌发病风险评估模型构建、模型验证分析 4 个方面，见图 1。

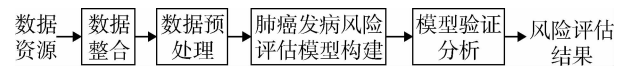


图 1 基于深度神经网络的老年肺癌发病风险评估原理

2.2.1 数据整合 梳理多年连续随访调查数据，将人口学基本特征（如性别、年龄、婚姻状况等）、生活习惯（如是否吸烟、是否喝酒等）、身体质量指数、吸烟情况（如目前吸烟情况、曾经吸烟情况等）、肺及气管相关疾病史（如慢性阻塞性肺疾病、哮喘等）、癌症病史情况、心脏相关疾病情况、是否糖尿病等多种危险因素整合到一起。根据受访者编号，将不同随访年份的各种危险因素一一对应，整合原始.xls 格式数据。例如，编号为 10013010 的受访者是否吸烟，在基线（第 0 年）时用“现在是否吸烟 0”表示，在首次随访（第 2 年）时用“现在是否吸烟 2”表示，以此类推，直到第 28 年末次随访时用“现在是否吸烟 28”表示，其他危险因素也用类似方法表示，将配偶出生年份、家中人数、子女数量、家庭资产收入情况等与本研究关联不大的变量舍去，整合后的数据样例，见表 1。经整合，共得到 18 396 条数据，涉及 56 个变量。

表 1 整合后数据样例

编号	出生时间（年）	性别	现在是否吸烟 0	现在是否吸烟 2	……	现在是否吸烟 28	是否患有癌症 0	……	是否患有癌症 28
10013010	1947	女	否	否	……	是	否	……	否
10281010	1938	女	否	否	……	否	否	……	否
12332030	1943	男	是	是	……	否	否	……	是
12496010	1941	男	否	否	……	否	否	……	否

2.2.2 数据预处理 对整合后的元数据进行一致性处理，并对危险因素指标的异常值进行校正和统一。由于随访时间较长，部分年份出现较多的空缺

值。采用多重插补法^[10]对空缺值进行填补。同时，通过奇异值分解方法^[11]降低数据中的白噪声，并按照受访者编码将多次随访数据进行关联匹配处理。

2.2.3 肺癌发病风险评估模型构建 采用深度神经网络方法构建老年肺癌发病风险评估模型。该模型通过计算不同危险因素的权重值，对肺癌发病风险进行定量评估。将预处理后的数据输入深度神经网络输入层，在训练过程中分别计算多种因素的权重值。权重值代表不同危险因素对肺癌是否发生的影响程度，通过设定权重阈值，模型最终输出老年肺癌的发病风险评估结果，具体流程，见图 2。

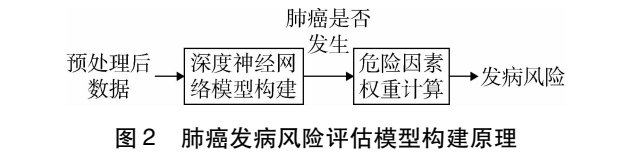


图2 肺癌发病风险评估模型构建原理

在隐藏层采用随机梯度下降法计算每层权重，训练过程中的激活函数使用修正线性单元（rectified linear unit, ReLU），如公式（1）所示。通过计算推导不同危险因素的权重值，如公式（2）所示，其中 W 是权重矩阵，变量 l 表示从第 2 层到第 L 层， b 是偏移向量，损失函数采用均方误差（mean - square error, MSE），如公式（3）所示。经过多次迭代，学习率按线性函数衰减，如公式（4）所示，其中 θ_0 为初始学习率， k 为衰减率， t 为训练迭代次数。采用 R 语言 4.3.3 和 H2O 框架，初始学习率为 0.01，批量大小为 32，经过多次学习训练得到隐藏层最优结构为 3 层，隐藏层有 22 个节点，输入层有 56 个节点，输出层有 1 个节点，在训练过程中使用十折交叉检验和网格搜索，隐藏层节点参数调整，见图 3。

$$\sigma(x)=\begin{cases}0,&x\leq0\\x,&x>0\end{cases}\tag{1}$$

$$\alpha^l=\sigma(W^l\alpha^{l-1}+b^l)\tag{2}$$

$$MSE=\frac{1}{L}\sum_{l=1}^L(\alpha^l-\alpha^{l'})^2\tag{3}$$

$$\theta=\theta_0-kt\tag{4}$$

同时求解权重阈值，见图 4。横轴上“0”表示随访基线第 0 年，以此类推。利用权重阈值对发病风险进行分层划分，权重均值大于等于阈值的划分为高风险，权重均值小于阈值的划分为低风险，进而输出对应的肺癌发病风险为高风险或低风险。按随访时间分别训练深度神经网络模型，得到不同年份肺癌发病风险评估结果。随机地将数据按照 8:2 分成训练集

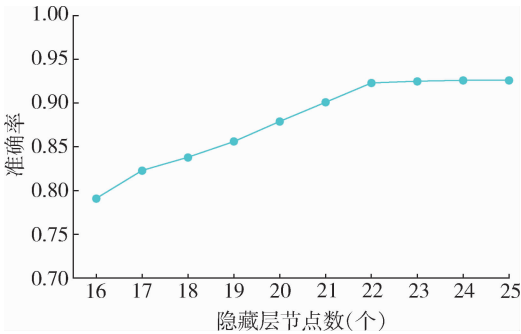


图3 隐藏层节点参数调整

和测试集，各自对应的模型准确率、曲线下面积（area under the curve, AUC），见图 5。为提高模型的可解释性，利用 Shapley 加法解释（Shapley additive explanations, SHAP）工具计算主要危险因素 SHAP 值，其对模型准确率提升的贡献度，见图 6。

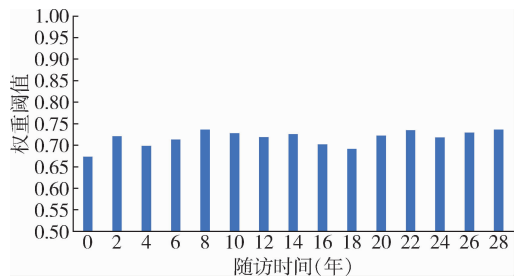


图4 肺癌发病风险评估权重阈值

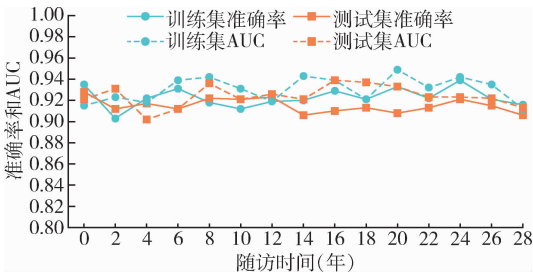


图5 肺癌发病风险评估模型准确率和 AUC

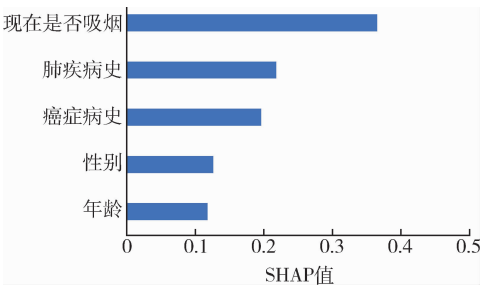


图6 主要危险因素 SHAP 值

2.2.4 模型验证分析 利用 Kaplan – Meier (KM) 算法进行老年肺癌生存分析, 比较不同发病风险的肺癌发病率和生存率, 以验证老年肺癌风险评估的有效性和可行性。同时, 与支持向量机、随机森林、逻辑回归等方法进行对比分析, 进一步验证基于深度神经网络方法的准确性。

3 结果

3.1 肺癌发病风险分析

利用肺癌发病风险评估模型得到老年肺癌发病高风险组和低风险组, 通过 KM 曲线展示测试集和训练集各组的肺癌发病生存率, 见图 7。

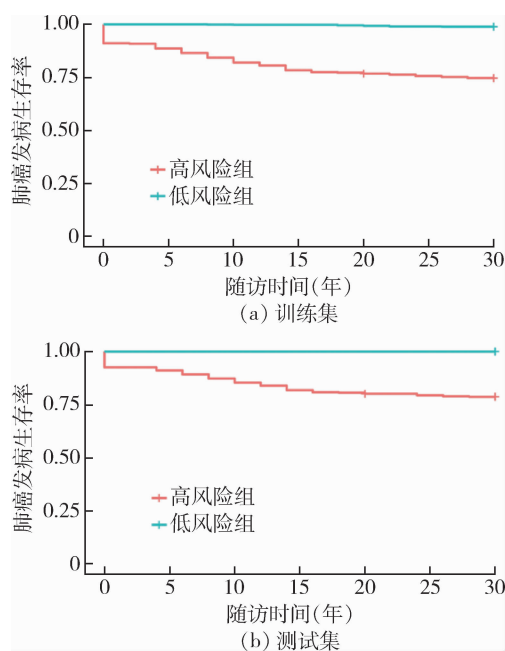


图 7 基于深度神经网络方法评估得到高风险组和低风险组的肺癌发病生存率

横轴表示随访时间, 纵轴“1.00”表示肺癌未发生, “0.75”表示肺癌发病生存率为 75%, 对应肺癌发病率为 25%。可见基于深度神经网络构建的老年肺癌发病风险评估模型得到的高风险组肺癌发病率明显高于低风险组。在末次随访时, 高风险组肺癌发病率约 25%, 远高于低风险组。同时, 通过训练集和测试集分别得到的高风险组肺癌发病率和低风险组肺癌发病率基本一致, 表明利用深度神经网络构建的老年肺癌发病风险评估模型具有可用性

和有效性。

不同风险组肺癌发病率, 见图 8。男性中高风险组肺癌发病率是低风险组的 6.7 倍, 女性中高风险组肺癌发病率是低风险组的 8 倍, 针对高风险组采取干预对有效控制老年肺癌发病有积极作用。且男性高风险组肺癌发病率是女性高风险组肺癌发病率的 2.02 倍, 在男性高风险组开展干预对降低肺癌发病率效果将更加显著。同时验证了基于深度神经网络开展肺癌发病风险评估研究的可行性。

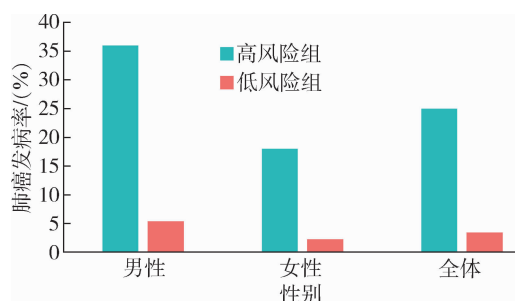


图 8 不同风险组肺癌发病率

3.2 模型验证分析

3.2.1 肺癌生存分析 利用 KM 曲线展示高风险组和低风险组肺癌生存率, 见图 9, 末次随访时低风险组的肺癌生存率较高, 高风险组肺癌生存率明显低于低风险组。结合前文图 7, 在多次随访中, 高风险组肺癌发病率明显高于低风险组, 且高风险组肺癌生存率均低于低风险组, 进一步验证了风险评估模型的有效性和可行性。

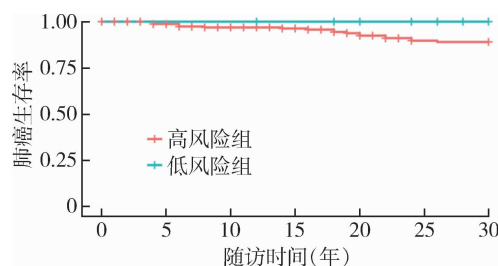


图 9 不同风险组肺癌生存率

3.2.2 对比分析 与支持向量机、随机森林、逻辑回归等方法对比, 基于深度神经网络方法构建的肺癌发病风险评估模型性能更优, 见图 10。

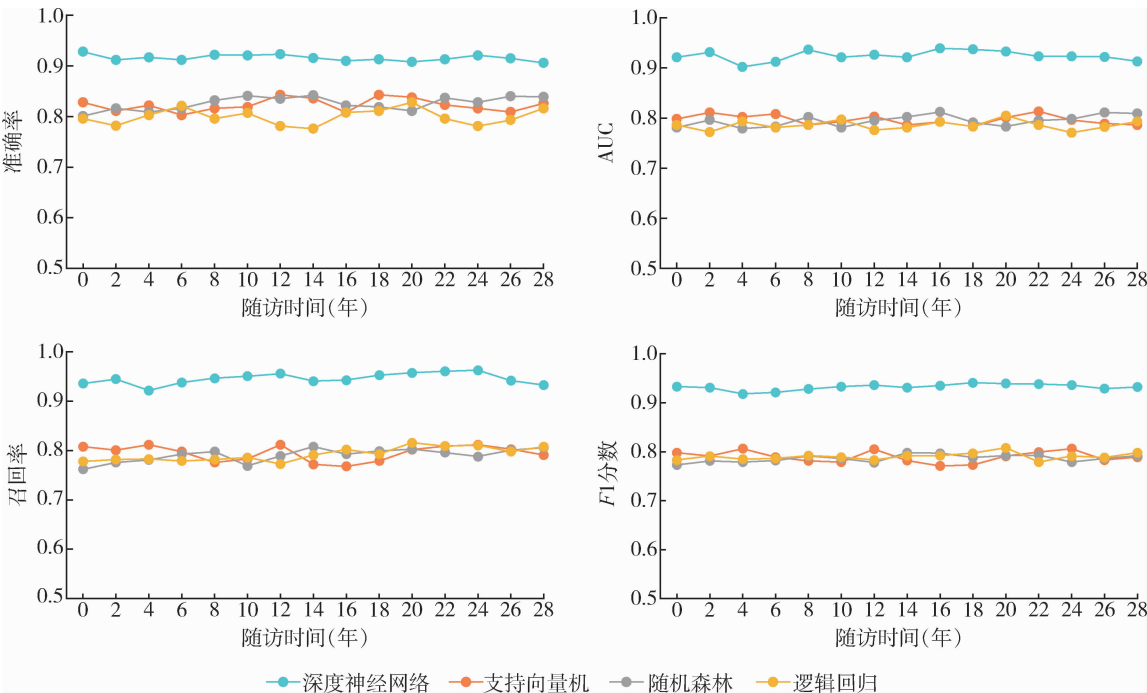


图 10 多种方法模型性能对比

4 讨论

本研究基于深度神经网络，采用大规模随访队列数据开展老年肺癌发病风险评估研究。通过验证和对比分析，显示了所建模型的有效性和准确性，为开展有针对性的肺癌发病干预控制提供了可靠依据，有利于老年肺癌的早诊早治。同时，本研究为提高老年人的生活质量和积极应对人口老龄化提供了有力保障。

近年来，肺癌发病风险评估相关研究大多采用传统统计方法^[12-13]或浅层机器学习方法^[14-16]，如支持向量机^[17]、随机森林^[18-19]、朴素贝叶斯^[20]、决策树算法^[17]等。Marzano L等^[21]利用随机森林和Cox回归分析，从小细胞肺癌的真实数据中改进了风险分层的预后亚组检测方法。Emaminejad N等^[20]采用朴素贝叶斯网络和多层感知机开展肺癌发病风险预测，以改善肺癌早期的预后评估。通过多元回归、独立t检验、单因素方差分析、皮尔逊相关系数等统计方法，对癌症健康素养、癌症知识、癌症预防行为、接受癌症筛查意愿等方面进行横截面在线问卷调查，以确定影响肺癌高危人群接受癌

症筛查的因素^[12]。基于决策树、支持向量机、随机森林、广义线性模型等研究改善非小细胞肺癌预后和促进个性化预防决策^[17]。哈佛大学研究团队^[22]从临床试验、评估筛查处理方法、戒烟干预措施、人工智能、生物标志物等方面对肺癌风险评估和筛查效率的提升开展系统分析。经过与常用的支持向量机、随机森林、逻辑回归等^[23-26]进行对比分析，本研究方法的准确率、AUC、召回率、F1分数等性能指标更高，进一步体现出深度神经网络在复杂数据处理和高维数据分析等方面的优越性，且该模型计算效率高，节省了大量运算资源 and 时间成本。本研究在临床实际应用中面临一些挑战和困难，如模型在不同种族、不同国家/地区人群中的适用性，临床数据采集、处理过程中的不一致性，以及如何应对数据异质性、噪声数据等方面需要进一步深入研究。未来，考虑将该模型封装成易用自动化风险评估工具，并与现有的临床诊断工具相结合，以提升其应用价值。

5 结语

积极开展老年肺癌发病风险评估，不仅能为肺

癌防控提供理论依据,还能为减轻疾病负担提供基础支撑。本研究存在一定局限性:一是医疗健康相关数据获取周期普遍较长,多年随访数据采集周期更长。此外,从原始采集到对外开放获取,公开资源需经历一系列环节,难免导致一些延迟,从而影响研究数据的时效性。二是针对大规模人群进行的连续多年长期随访,时间跨度大,会面临诸多不可抗因素,可能导致数据缺失、产生白噪声等问题,给数据处理造成不少困难,并在一定程度上影响模型性能的提升。未来考虑将该模型融入电子健康档案等时序信息中,不仅可提高数据获取时效性,还能实时建模,优化并提升模型性能。同时,这也有助于实现适应我国人群的肺癌自动化风险评估,逐步深化本研究在重大疾病预防控制中的有益效果。进一步加强肺癌高风险人群的评估识别和预防干预,将有利于医疗资源的高效配置和疾病早诊早治,且有助于提升公众的疾病预防健康素养和全生命周期的健康水平。

作者贡献: 陈松景负责数据采集、整理与分析,论文撰写与修订;吴思竹负责论文审核与修订。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

- 1 ABERLE D R, BLACK W C, CHILES C, et al. Lung cancer incidence and mortality with extended follow-up in the national lung screening trial [J]. *Journal of thoracic oncology*, 2019, 14 (10): 1732-1742.
- 2 HOREWEG N, SCHOLTEN E T, DE JONG P A, et al. Detection of lung cancer through low-dose CT screening: a prespecified analysis of screening test performance and interval cancers [J]. *Lancet oncology*, 2014, 15 (12): 1342-1350.
- 3 NEMLANDER E, ROSENBLAD A, ABEDI E, et al. Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, former smokers and current smokers [J]. *Plos one*, 2023, 17 (10): e0276703.
- 4 ADAMEK M, WACHULA E, SZABLOWSKA S, et al. Risk factors assessment and risk prediction models in lung cancer screening candidates [J]. *Annals of translational medicine*, 2016, 4 (8): 151.
- 5 PETIT P, MAITRE A, PERSOONS R, et al. Lung cancer

risk assessment for workers exposed to polycyclic aromatic hydrocarbons in various industries [J]. *Environment international*, 2019, 124 (3): 109-120.

- 6 HE S Y, XIA C F, LI H, et al. Cancer profiles in China and comparisons with the USA: a comprehensive analysis in the incidence, mortality, survival, staging, and attribution to risk factors [J]. *Science China-life sciences*, 2024, 67 (1): 122-131.
- 7 GRANT A, MIN B, ELMRAYED S, et al. Prevalence, risk factors, and outcomes of adult interstitial lung abnormalities a systematic review and meta-analysis [J]. *American journal of respiratory and critical care medicine*, 2023, 208 (6): 695-708.
- 8 KE L H, CUI S P, YANG M, et al. Validation of a modified Caprini risk assessment model in lung cancer patients undergoing surgery: results of a multicenter cross-sectional observational study [J]. *Journal of surgical oncology*, 2022, 125 (5): 933-942.
- 9 KLEIN AWERJANOW K, RZYMAN W, DZIEDZIC R, et al. Assessment of calcium score cutoff point for clinically significant aortic stenosis on lung cancer screening program low-dose computed tomography - a cross-sectional analysis [J]. *Diagnostics*, 2023, 13 (2): 246.
- 10 SCHAFER J. Multiple imputation: a primer [J]. *Statistical methods in medical research*, 1999, 8 (1): 3-15.
- 11 KALMAN D. Singularly valuable decomposition: the SVD of a matrix [J]. *The college mathematics journal*, 1996, 27 (1): 1-23.
- 12 CHO M K, CHO Y H. Factors influencing the intention for lung cancer screening in high-risk populations for lung cancer [J]. *Asia-Pacific journal of oncology nursing*, 2024, 11 (1): 100332.
- 13 GE C, PETERS S, OLSSON A, et al. Diesel engine exhaust exposure, smoking, and lung cancer subtype risks a pooled exposure-response analysis of 14 case-control studies [J]. *American journal of respiratory and critical care medicine*, 2020, 202 (3): 402-411.
- 14 HSU Y C, TSAI Y H, WENG H H, et al. Artificial neural networks improve LDCT lung cancer screening: a comparative validation study [J]. *BMC cancer*, 2020, 20 (1): 1023.
- 15 WANG K M, CHEN K H, HERNANDA C A, et al. How is the lung cancer incidence rate associated with environmental risks? Machine-learning-based modeling and benchmark-

- ing [J]. International journal of environmental research and public health, 2022, 19 (14): 8445.
- 16 CHANG G C, CHIU C H, YU C J, et al. Low - dose CT screening among never - smokers with or without a family history of lung cancer in Taiwan: a prospective cohort study [J]. Lancet respiratory medicine, 2024, 12 (2): 141 - 152.
- 17 TANG F H, FONG Y W, YUNG S H, et al. Radiomics - clinical AI model with probability weighted strategy for prognosis prediction in non - small cell lung cancer [J]. Bio-medicines, 2023, 11 (8): 2093.
- 18 LEVI M, LAZEBNIK T, KUSHNIR S, et al. Machine learning computational model to predict lung cancer using e-lectronic medical records [J]. Cancer epidemiology, 2024, 92 (10): 102631.
- 19 ALTUHAIFA F A, WIN K T, SU G X. Predicting lung cancer survival based on clinical data using machine learning: a review [J]. Computers in biology and medicine, 2023, 165 (10): 107338.
- 20 EMAMINEJAD N, QIAN W, GUAN Y B, et al. Fusion of quantitative image and genomic biomarkers to improve prognosis assessment of early stage lung cancer patients [J]. IEEE transactions on biomedical engineering, 2016, 63 (5): 1034 - 1043.
- 21 MARZANO L, DARWICH A S, TENDLER S, et al. A novel analytical framework for risk stratification of real - world data using machine learning: a small cell lung cancer study [J]. CTS - clinical and translational science, 2022, 15 (10): 2437 - 2447.
- 22 ADAMS S J, STONE E, BALDWIN D R, et al. Lung cancer screening [J]. Lancet, 2023, 401 (10374): 390 - 408.
- 23 王兰荣, 王晓翠, 曹畅, 等. 郑州市肺癌早期筛查患病风险横断面研究 [J]. 实用医学杂志, 2024, 40 (15): 2154 - 2160.
- 24 NUNEZ E R, ZHANG S Q, GLICKMAN M E, et al. What goes into patient selection for lung cancer screening? Factors associated with clinician judgments of suitability for screening [J]. American journal of respiratory and critical care medicine, 2024, 209 (2): 197 - 205.
- 25 孟诚, 邬芝雅, 冯继锋. 肺结节风险分层评估的研究进展 [J]. 中国现代医生, 2023, 61 (26): 141 - 144.
- 26 MCDOWELL A, KANG J, YANG J, et al. Machine - learning algorithms for asthma, COPD, and lung cancer risk assessment using circulating microbial extracellular vesicle data and their application to assess dietary effects [J]. Experimental and molecular medicine, 2022, 54 (9): 1586 - 1595.

(上接第 15 页)

- 34 CHOWDHARY N, JOTHEESWARAN A T, NADKARNI A, et al. The methods and outcomes of cultural adaptations of psychological treatments for depressive disorders: a systematic review [J]. Psychological medicine, 2014, 44 (6): 1131 - 1146.
- 35 LARANJO L, DUNN A G, TONG H L, et al. Conversational agents in healthcare: a systematic review [J]. Journal of the American medical informatics association, 2018, 25 (9): 1248 - 1258.
- 36 AYERS J W, POLIAK A, DREDZE M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum [J]. JAMA internal medicine, 2023, 183 (6): 589 - 596.
- 37 GOTZL C, HILLER S, RAUSCHENBERG C, et al. Artificial intelligence - informed mobile mental health apps for young people: a mixed - methods approach on user' and stakeholders' perspectives [J]. Child and adolescent psychiatry and mental health, 2022, 16 (1): 86.
- 38 TRAN S, SMITH L, EL - DEN S, et al. The use of gamification and incentives in mobile health apps to improve medication adherence: scoping review [J]. JMIR mhealth and uhealth, 2022, 10 (2): e30671.
- 39 CHENG V, DAVENPORT T, JOHNSON D, et al. Gamification in apps and technologies for improving mental health and well - being: systematic review [J]. JMIR mental health, 2019, 6 (6): e13717.
- 40 TOPOL E J. High - performance medicine: the convergence of human and artificial intelligence [J]. Nature medicine, 2019, 25 (1): 44 - 56.