

· 论 著 · DOI:10.3969/j.issn.1672-9455.2025.04.010

运用 Lasso 回归模型预测重庆南川地区呼吸道感染的研究*

马 臣¹, 邹进美², 夏孟红^{3△}

重庆市南川区人民医院:1. 医保办;2. 病案管理科;3. 骨科, 重庆 408400

摘要:目的 探讨温度变化和空气质量对呼吸道感染的影响,为预防呼吸道感染提供帮助。方法 获取 2021 年 1 月 1 日至 2023 年 2 月 28 日于该院门诊因呼吸道感染就诊的 160 521 例患者的就诊信息以及同期该地区温度、4 类大气污染物浓度数据,按照就诊日期将 2021 年 1 月至 2022 年 12 月每日发生病例数及温度、4 类大气污染物浓度数据作为训练集,将 2023 年 1—2 月的相关数据作为测试集。利用 Lasso 回归在训练集中选择影响呼吸道感染发生的因子,建立预测模型,并对 2023 年 1—2 月(59 d)的感染人数进行预测,与测试集中实际发生呼吸道感染的人数进行比较。结果 基于 Lasso 回归,影响该地区呼吸道感染发生的主要因素有 6 个:当日 PM_{2.5} 浓度(X_1)、前 3 d PM_{2.5} 浓度(X_2)、当日二氧化硫(SO₂)浓度(X_7)、当日气温(X_9)、当日温差(X_{11})和前 3 d 温差(X_{12}),预测模型为 $Y=250.64+2.90X_1+0.90X_2+0.75X_7-4.84X_9+4.09X_{11}+1.29X_{12}$ 。预测值与实际值平均绝对百分数误差为 0.21,模型预测精度较高,前 10 d(2023 年 1 月 1—10 日)平均绝对百分数误差为 0.07,短期预测精度较高。结论 低温对呼吸道感染影响最大,PM_{2.5} 浓度和温差对呼吸道感染存在滞后效应;该预测模型更适用于短期预测。

关键词:气候变化; 空气质量; 呼吸道感染; Lasso 回归; 预测

中图法分类号:R183.3;R562

文献标志码:A

文章编号:1672-9455(2025)04-0480-05

Study on the prediction of respiratory tract infections in Nanchuan area of Chongqing by using the Lasso regression model*MA Chen¹, ZOU Jinmei², XIA Menghong^{3△}

1. Department of Medical Insurance Office; 2. Department of Medical Records Management; 3. Department of Orthopaedics, the People's Hospital of Nanchuan, Chongqing 408400, China

Abstract: Objective To investigate the impact of temperature changes and air quality on respiratory tract infections, and to provide assistance for the prevention of respiratory tract infections. **Methods** Data on 160 521 patients who visited the outpatient department of the hospital for respiratory tract infections from January 1, 2021 to February 28, 2023 were collected, along with corresponding temperature data and concentrations of four types of air pollutants. Daily case numbers, temperature and concentrations of the four pollutants from January 2021 to December 2022 were used as the training set, while the relevant data from January to February 2023 were used as the testing set. Lasso regression was employed on the training dataset to identify key factors influencing the occurrence of respiratory tract infections and to establish a predictive model. The model was used to forecast the number of respiratory tract infections over a 59-day period from January to February 2023. Predicted infection counts were compared with the actual number of respiratory tract infections recorded in the test dataset during the same period. **Results** Based on Lasso regression, six main factors influencing respiratory tract infections in the region were identified: the PM_{2.5} concentration on the day (X_1), the PM_{2.5} concentration over the previous 3 days (X_2), the sulfur dioxide (SO₂) concentration on the day (X_7), the temperature on the day (X_9), the daily temperature difference (X_{11}) and the temperature difference over the previous 3 days (X_{12}). The predictive model was established as: $Y=250.64+2.90X_1+0.90X_2+0.75X_7-4.84X_9+4.09X_{11}+1.29X_{12}$. The mean absolute percentage error (MAPE) between the predicted and actual values was 0.21, indicating high model prediction accuracy. The MAPE for the first 10 d (January 1 to 10,

* 基金项目:重庆市科卫联合医学科研项目(2022QNXM044);重庆市自然科学基金面上项目(cstc2021jcyj-msxmX0700)。

作者简介:马臣,男,统计师,主要从事医院管理相关研究。

2023) was 0.07, demonstrating high short-term prediction accuracy. **Conclusion** Low temperature has the greatest impact on respiratory tract infections, while PM_{2.5} concentration and temperature difference exhibit delayed effects on respiratory tract infections. The predictive model is more suitable for short-term forecasting.

Key words: climate change; air quality; respiratory tract infection; Lasso regression; prediction

呼吸道感染作为人类最常见的传染病,已成为全球,特别是发展中国家婴幼儿和老年人住院的主要原因^[1]。据统计,在发展中国家中因急性上呼吸道感染导致的死亡约占 75%^[2],在我国农村居民和城市居民死亡病因中呼吸道感染排名均靠前^[3]。目前,国内外学者在不同地区开展气象因素对呼吸道感染的影响评估,均发现气温变化与气候条件对感染发生具有显著影响^[4-5]。及时探寻发生呼吸道感染的危险因素,制订个体化防治策略,可降低呼吸道感染的发病率和传播率^[6]。Lasso 回归模型在多维影响因子筛选中应用较为广泛^[7],与传统回归模型比较,其在变量选择上更稳健和准确,可以有效避免多重共线性对模型的干扰,提高模型的预测精度,其基本思想是在最小二乘法基础上加上一个范数正则项,使系数不断被压缩至 0,从高维变量中高效提取出关键变量^[8]。基于此,本研究拟从医院门诊数据出发,利用 Lasso 变量选择方法探讨气温变化和空气质量状况与呼吸道感染发生的关系,并建立 Lasso 回归模型预测呼吸道感染发生人数,以供医院备足医疗资源。

1 资料与方法

1.1 一般资料 获取 2021 年 1 月 1 日至 2023 年 2 月 28 日重庆市南川区人民医院门诊患者信息。纳入标准:(1)门诊主要诊断编码符合国际疾病分类标准第 10 版(ICD-10)中编码为 J00~J99 的病例;(2)符合《急性上呼吸道感染基层诊疗指南(2018 年)》^[9]中呼吸道感染诊断标准;(3)临床表现均为咽痛、咳嗽、肺部炎症等。排除标准:由癌性发热、术后并发症等其他疾病引发的呼吸道感染。将 2021 年 1 月 1 日至 2022 年 12 月 31 日每日感染人数和同期温度、大气污染物浓度作为训练集样本,平均每日感染人数 204 例(合计 149 609 例),男女比例约为 110:100,年龄 3~52 岁、中位年龄 25 岁。将 2023 年 1 月 1 日至 2023 年 2 月 28 日相关数据作为测试集样本,平均每日感染人数 185 例(合计 10 912 例),男女比例约为 112:100,年龄 4~55 岁、中位年龄 31 岁。本研究已获得本院医学伦理委员会审批(YXYJ-2021-009-01)。

1.2 变量选取 世界气象组织(WMO)权威专家强调,全球气候变暖趋势仍在继续,气温作为气候变化的主要指标之一,呼吸道感染受其改变影响显著^[10],结合临床经验,温度低、温差变化大时,呼吸道感染人

数多。因此,本研究考虑将平均气温及温度差作为自变量,分析其与呼吸道感染发生的数量关系。此外,研究表明,大气污染物中 PM_{2.5}、PM₁₀、二氧化硫(SO₂)、二氧化氮(NO₂)浓度对呼吸道感染具有不同程度的影响^[11],因此,本研究考虑将这 4 种具有代表性的污染物浓度作为自变量。结合呼吸道感染的发生存在 1~3 d 的潜伏期^[6],变量选取时考虑纳入温度和污染物浓度提前 3 d 的数据,探讨对发病的影响。鉴于此,本研究将前 3 d 温度差、当日温度差、前 3 d 平均气温、当日平均气温、前 3 d PM_{2.5} 浓度、当日 PM_{2.5} 浓度、前 3 d PM₁₀ 浓度、当日 PM₁₀ 浓度、前 3 d SO₂ 浓度、当日 SO₂ 浓度、前 3 d NO₂ 浓度、当日 NO₂ 浓度共 12 个变量作为自变量,研究其对上呼吸道感染人数的影响,建立预测模型。

1.3 统计学处理 采用 SPSS22.0 统计软件进行数据分析,符合正态分布计量资料以 $\bar{x} \pm s$ 表示,组间比较采用 *t* 检验;不符合正态分布的计量资料采用 *M*(*P*₂₅, *P*₇₅)表示,组间比较采用秩和检验;计数资料以例数和百分率表示,组间比较采用 χ^2 检验。变量间的相关性采用 Pearson 相关进行分析。采用 R4.3.2 统计软件进行模型建立,以最优 λ 值(λ 最小时)筛选变量。利用平均绝对百分数误差验证预测值与实际值的差异。以 *P* < 0.05 为差异有统计学意义。

2 结果

2.1 患者一般情况 共纳入 160 521 例患者:男 84 169 例(52.4%),女 76 352(47.6%);0~6 岁 55 292 例,>6~18 岁 17 035 例,>18~60 岁 60 399 例,>60 岁 27 795 例。将 160 521 份数据资料按照时间先后顺序分为训练集(149 609 例)和测试集(10 912 例)。训练集和测试集患者的性别和年龄构成比较,差异均无统计学意义(*P* > 0.05)。见表 1。

表 1 呼吸道感染人群性别与年龄构成[n(%)]

项目	训练集 (n=149 609)	测试集 (n=10 912)	χ^2	<i>P</i>
性别			1.079	0.299
男	78 395(52.4)	5 774(52.9)		
女	71 214(47.6)	5 138(47.1)		
年龄			2.987	0.394
0~6 岁	51 585(34.5)	3 707(34.0)		

续表 1 呼吸道感染人群性别与年龄构成[n(%)]

项目	训练集 (n=149 609)	测试集 (n=10 912)	χ^2	P
>6~18 岁	15 888(10.6)	1 147(10.5)		
>18~60 岁	56 209(37.6)	4 190(38.4)		
>60 岁	25 927(17.3)	1 868(17.1)		

2.2 变量特征 2021 年 1 月 1 日至 2022 年 12 月 31 日该院平均每日呼吸道感染发生中位人数为 175 例。温度、温差、空气污染物分布特征见表 2。定义人数为 Y, 当日 PM_{2.5} 浓度为 X₁、前 3 d PM_{2.5} 浓度为 X₂、当日 PM₁₀ 浓度为 X₃、前 3 d PM₁₀ 浓度为 X₄、当日 NO₂ 浓度为 X₅、前 3 d NO₂ 浓度为 X₆、当日 SO₂ 浓度为 X₇、前 3 d SO₂ 浓度为 X₈、当日平均气温为 X₉、前 3 d 平均气温为 X₁₀、当日温差为 X₁₁、前 3 d 温差为 X₁₂。

表 2 温度、温差、空气污染物分布特征

变量	\bar{x}	s	Min	P ₂₅	M	P ₇₅	Max
当日 PM _{2.5} (μg/m ³)	33.13	21.0	7	17	26	43	117
前 3 d PM _{2.5} (μg/m ³)	33.15	21.0	7	17	26	43.25	117
当日 PM ₁₀ (μg/m ³)	50.88	27.14	12	30	44	65	159
前 3 d PM ₁₀ (μg/m ³)	50.91	27.14	12	30	44	65	159
当日 NO ₂ (μg/m ³)	30.03	11.28	4	22	28	37	86
前 3 d NO ₂ (μg/m ³)	30.04	11.29	4	22	28	37	86
当日 SO ₂ (μg/m ³)	10.37	6.28	6	8	9	11	97
前 3 d SO ₂ (μg/m ³)	10.37	6.29	6	8	9	11	97
当日平均气温(°C)	21.43	9.28	2	14	21	29	40
前 3 d 平均气温(°C)	21.43	9.29	2	14	21	29	40
当日温差(°C)	6.84	3.67	1	4	6	10	16
前 3 d 温差(°C)	6.84	3.67	1	4	6	10	16

2.3 温度、温差、空气污染物与感染人数的相关性分析 Pearson 相关分析结果显示, 呼吸道感染日发生人数与温度、空气污染物之间存在相关关系(P < 0.01)。其中, 呼吸道感染日发生人数与当日 PM_{2.5} 浓度、前 3 d PM_{2.5} 浓度、当日 PM₁₀ 浓度、前 3 d PM₁₀ 浓度、当日 NO₂ 浓度、前 3 d NO₂ 浓度、当日 SO₂ 浓度、前 3 d SO₂ 浓度的相关系数 r 分别为 0.700、0.590、0.600、0.560、0.415、0.342、0.274、0.215, 呼吸道感染日发生人数与当日平均气温、前 3 d 平均气温、当日温差、前 3 d 温差的相关系数 r 分别为 -0.543、-0.552、-0.163、-0.139。此外, 不同自变量间具有显著相关性, 说明变量间存在严重的多重共线性, 因此有必要对变量进行选择。

2.4 呼吸道感染影响因素的筛选 为消除变量之间

的多重共线性, 提高模型预测精度, 采用 Lasso 回归对变量进行筛选, 随着参数 log(λ) 增大, 回归系数不断压缩至 0(图 1), 依据 lambda.min 变量筛选准则, 交叉验证结果显示最优结果对应 λ=1.48, 共筛选出系数不为 0 的变量 6 个, 包括当日 PM_{2.5} 浓度、前 3 d PM_{2.5} 浓度、当日 SO₂ 浓度、当日平均气温、当日温差和前 3 d 温差, 系数分别为 2.90、0.90、0.75、-4.84、4.09、1.29, 得到预测模型 Y=250.64+2.90X₁+0.90X₂+0.75X₇-4.84X₉+4.09X₁₁+1.29X₁₂。见图 2。

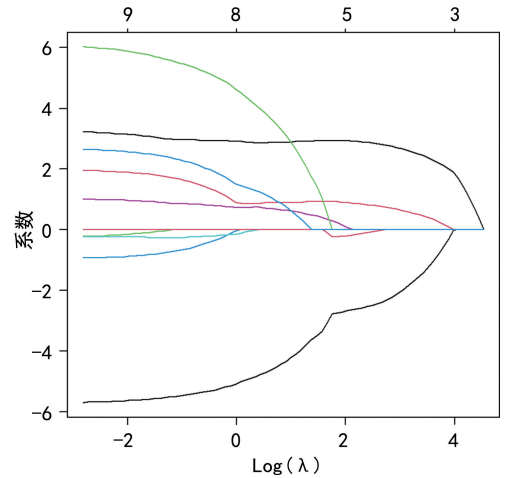


图 1 系数分布图

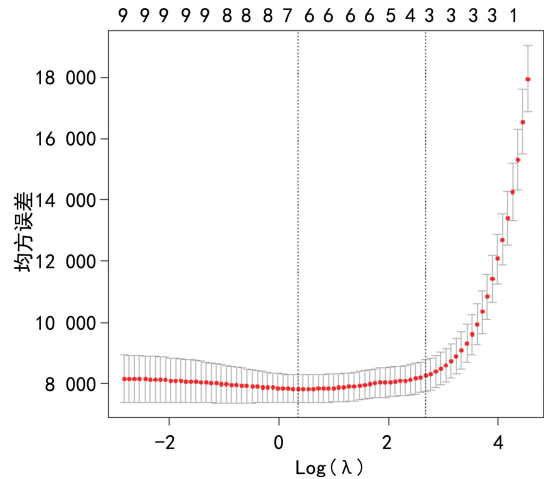


图 2 LASSO 回归交叉验证曲线图

2.5 呼吸道感染发生人数的预测 将测试集中当日 PM_{2.5} 浓度、前 3 d PM_{2.5} 浓度、当日 SO₂ 浓度、当日平均气温、当日温差和前 3 d 温差代入上述模型, 对每日呼吸道感染发生人数进行预测, 得到的预测结果与测试集中实际发生人数进行比较, 结果见图 3。该模型平均绝对百分数误差为 0.21, 预测精度较高, 其中前 10 d(2023 年 1 月 1—10 日)的预测值与实际值相对误差控制在 10% 以内, 平均绝对百分数误差为 0.07。随着预测天数增加, 预测结果误差越大, 但预测值与实测值走势相同。

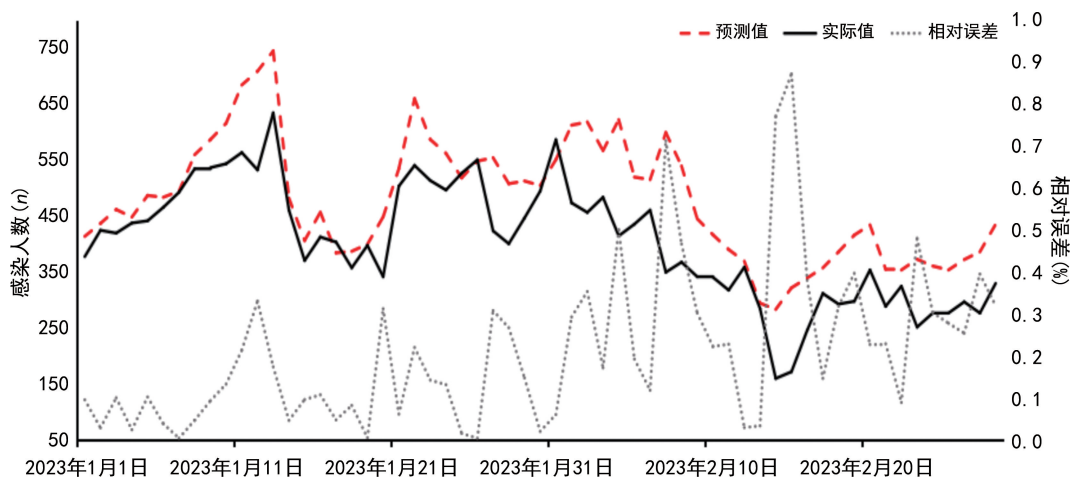


图 3 感染人数预测值与实际值比较

3 讨论

本研究利用门诊数据分析重庆市南川地区呼吸道感染发生的特征,探讨感染发生人数与温度变化、空气质量的关系,进而对感染人数进行预测,为呼吸道感染的预防提供依据。主要得到以下结论。

3.1 PM_{2.5} 对呼吸道感染的影响 当日 PM_{2.5} 系数为 2.90,该值表示在保持其他条件不变的情况下,空气中 PM_{2.5} 浓度越高,呼吸道感染发生人数越多。PM_{2.5} 对呼吸系统的致病机制为其作为空气中悬浮的微小颗粒物,附着多种病原体进入人体呼吸道,攻击呼吸系统^[12]。其次,PM_{2.5} 会促进氧化应激和血小板活化因子,诱导细胞衰老,降低气道抗菌蛋白肽的表达,使病原体黏附于气道上皮细胞,导致呼吸道的抗菌活性降低,增加感染风险^[13]。前 3 天 PM_{2.5} 系数为 0.90,表明 PM_{2.5} 对感染人数的预测存在滞后影响。因此,在 PM_{2.5} 浓度持续上升的情况下,疾病预防控制部门需要提高警惕,控制呼吸道感染蔓延。

3.2 SO₂ 浓度对呼吸感染的影响 当日 SO₂ 浓度系数为 0.75,表示在保持其他条件不变的情况下,SO₂ 浓度越高,呼吸道感染发生人数越多。SO₂ 是一种无色有毒气体,其刺激性成分进入呼吸道会损伤黏膜,从而出现胸闷气短、呼吸疼痛等情况。

3.3 气温对呼吸道感染的影响 当日平均气温系数为 -4.84,表示在其他条件保持不变的情况下,气温越低,呼吸道感染发生人数越多,且该地区温度变化对呼吸道感染发生的影响最大,气温降低,冷空气突然来袭,对呼吸道产生刺激,使鼻咽部局部黏膜变得干燥,从而导致呼吸道黏膜抵抗感染的能力受损,病毒、细菌更易乘虚而入,从而引发呼吸道感染^[14]。在国外,如泰国、韩国均已报道低温更易引发呼吸道感染^[15-16],同样在我国大部分地区均是低温天气易引发呼吸道感染,如淮安市低温容易引起呼吸道感染日门诊量增加^[17],深圳地区冬、春季冷空气活动,以及其所

反映的低温、低湿、大风等产生的冷效应是诱发上呼吸道感染的关键因素^[18]。此外,当日温差系数为 4.09,表示温差越大呼吸道感染更容易发生;前 3 d 温差系数为 1.29,表示温差对呼吸道感染发生存在滞后影响,在我国浙江湖州、辽宁北票地区也证实了呼吸道感染发生与日气温的相关性显著,且存在滞后效应^[19-20]。低温天气是病原菌感染好发季节,同时剧烈的温度变化会导致人体抵抗能力减弱,因此更容易引发呼吸道感染,在低温环境下,需要加强身体锻炼,注意防寒保暖。

3.4 模型预测效果 对呼吸道感染的监测是公共卫生管理的一项重要工作,因此,本研究利用 Lasso 回归模型对呼吸道感染人数进行预测,并与测试集中实际发生人数进行比较,整体预测趋势与实际数据走势符合,说明 Lasso 回归选取的指标具有代表性,可为当地防污治理提供精准抓手。此外,模型预测前 10 d 的相对误差控制在 10% 以内,平均绝对百分数误差为 0.07,表明该模型在呼吸道感染人数短期预测中精度较高,可为医疗机构准备医疗资源提供合理数据,避免因物资不足影响患者就医或购药,为公共卫生决策提供支撑。

参考文献

- [1] XING X F, TANG S Q, WANG W Y, et al. Trends in mortality from lower respiratory infections in children under 14 years, 1990—2019: an analysis of the global burden of disease study 2019[J]. *Heliyon*, 2024, 10(21): e39502.
- [2] ZIMMERMAN R K, RINALDO C R, NOWALK M P, et al. Viral infections in outpatients with medically attended acute respiratory illness during the 2012—2013 influenza season[J]. *BMC Infect Dis*, 2015, 15: 87.
- [3] 国家统计局. 2021 年我国部分地区城市居民主要疾病死亡率及死因构成[J]. *中国病毒病杂志*, 2023, 13(6): 476.
- [4] SCHWARZ M, PETERS A, STAFIOGGIA M, et al. Tem-

- poral variations in the short-term effects of ambient air pollution on cardiovascular and respiratory mortality: a pooled analysis of 380 urban areas over a 22-year period [J]. *Lancet Planetary Health*, 2024, 8(9): e657-e665.
- [5] 郑甲炜, 王式功, 尹立, 等. 基于机器学习的两个代表城市上呼吸道感染与气象要素关系及其发病风险预测研究[J]. *沙漠与绿洲气象*, 2023, 17(4): 160-168.
- [6] 刘波, 古贞. 2016—2021年四川自贡地区急性上呼吸道感染与气象因素关联分析[J]. *公共卫生与预防医学*, 2022, 33(4): 37-40.
- [7] BODINIER B, FILIPPI S, NOST T H, et al. Automated calibration for stability selection in penalised regression and graphical models[J]. *J R Stat Soc Ser C Appl Stat*, 2023, 72(5): 1375-1393.
- [8] 赵向荣. 甘肃劳动者报酬影响因素分析: 基于岭回归与 Lasso 回归[J]. *商展经济*, 2024, 6(6): 118-121.
- [9] 陈云天, 熊维宁. 《急性上呼吸道感染基层诊疗指南(2018年)》的解读[J]. *中华全科医师杂志*, 2020, 19(1): 86-88.
- [10] ZHANG X S, XU K, LI Q, et al. Exploring the impact of ambient temperature on respiratory diseases admissions, length of stay, and hospitalization costs in Lanzhou City, based on distributed lag non-linear model [J]. *Climate Services*, 2024, 34: 100481.
- [11] 赵容, 洪也, 郑国峰, 等. 沈阳市大气污染物 PM_{2.5}、PM₁₀、SO₂、NO₂ 浓度对呼吸系统疾病严重程度的影响[J]. *山东医药*, 2022, 62(23): 11-14.
- [12] 易建华, 吴晓芳, 王丽云, 等. PM_{2.5} 对呼吸系统疾病的影响及其机制的研究进展[J]. *西安交通大学学报(医学版)*, 2019, 40(1): 167-172.
- [13] ZHANG S C, HUO X, ZHANG Y, et al. Ambient fine particulate matter inhibits innate airway antimicrobial activity in preschool children in e-waste areas [J]. *Environ Int*, 2019, 123: 535-542.
- [14] ECCLES R. An explanation for the seasonality of acute upper respiratory tract viral infections [J]. *Acta Otolaryngol*, 2002, 122(2): 183-191.
- [15] THONGPAN I, VONGPUNSAWAD S, POOVORAWAN Y. Respiratory syncytial virus infection trend is associated with meteorological factors [J]. *Sci Rep*, 2020, 10(1): 10931.
- [16] LEE H, YOON H Y. Impact of ambient temperature on respiratory disease: a case-crossover study in Seoul [J]. *Respir Res*, 2024, 25(1): 73.
- [17] 徐梦颖, 骆善彩, 谢真珍, 等. 气象因素对淮安市呼吸系统疾病日门诊量的影响[J]. *江苏预防医学*, 2023, 34(2): 156-159.
- [18] 黄开龙, 林锦春, 马盼, 等. 气象条件对深圳市罗湖区上呼吸道感染就诊人数的影响[J]. *干旱气象*, 2021, 39(6): 995-1005.
- [19] 徐澳, 刘光涛, 刘艳, 等. 基于分布滞后非线性模型的环境因素与湖州市流感样病例发病的关联及滞后性研究[J]. *中华疾病控制杂志*, 2023, 27(7): 783-787.
- [20] 殷宁璐, 李俊林, 洪也, 等. 辽宁省两县域城市气温对呼吸系统疾病住院人数的影响[J]. *干旱气象*, 2023, 41(1): 132-142.
- (收稿日期: 2024-07-12 修回日期: 2024-11-18)
-
- (上接第 479 页)
- [16] 王蕾, 邱胜利, 曹俊, 等. 脑卒中后肺部多药耐药菌感染危险因素及预测模型构建[J]. *中华医院感染学杂志*, 2022, 32(7): 990-993.
- [17] ALIBERTI S, COOK G S, BABU B L, et al. International prevalence and risk factors evaluation for drug-resistant *Streptococcus pneumoniae* pneumonia [J]. *J Infect*, 2019, 79(4): 300-311.
- [18] MECHERGUI A, ACHOUR W, MATHLOUTHI S, et al. Prevalence of infectious multi-drug resistant bacteria isolated from immunocompromised patients in Tunisia [J]. *Afr Health Sci*, 2019, 19(2): 2021-2025.
- [19] SON H J, KIM T, LEE E, et al. Risk factors for isolation of multi-drug resistant organisms in coronavirus disease 2019 pneumonia: a multicenter study [J]. *Am J Infect Control*, 2021, 49(10): 1256-1261.
- [20] BAIDYA S, SHARMA S, MISHRA S K, et al. Biofilm formation by pathogens causing ventilator-associated pneumonia at intensive care units in a tertiary care hospital: an armor for refuge [J]. *Biomed Res Int*, 2021, 2021: 8817700.
- [21] 王秋义, 尤丕聪. 脑卒中相关性肺炎病原菌特点及风险列线图模型建立[J]. *中华医院感染学杂志*, 2021, 31(7): 967-971.
- [22] MIRTALAEI N, FARAZI A, EBRAHIMI MONFARED M, et al. Efficacy of antibiotic prophylaxis against ventilator-associated pneumonia [J]. *J Hosp Infect*, 2019, 101(3): 272-275.
- [23] RASHID M H, KABIR A, WARIS M U, et al. Role of prophylactic antibiotics in critical care of stroke patients: a preventive approach to post-stroke infections [J]. *Cureus*, 2020, 12(3): e7158.
- (收稿日期: 2024-07-12 修回日期: 2024-10-18)