

基于机器学习的儿科急诊患者候诊时间预测及关联分析研究^{*}

郭琳琳 魏永祥 郭琳瑛 李 姣 郑 思

(首都儿科研究所附属儿童医院 北京 100020) (中国医学科学院/北京协和医学院医学信息研究所 北京 100020)

崔 英 陈 宇

(首都儿科研究所附属儿童医院 北京 100020)

〔摘要〕 以首都儿科研究所附属儿童医院急诊患者就诊数据为基础, 设计并提取 19 个就诊时间相关变量, 构建 6 种不同的机器学习模型, 用于儿科急诊就诊患者候诊时间预测, 经检验预测模型表现良好。本研究有助于医院动态分配医疗资源, 缓解急诊科拥挤程度, 提高患者就诊满意度。

〔关键词〕 儿科急诊; 机器学习; 候诊时间; 预测

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2022.04.006

Study on Prediction and Correlation Analysis of Pediatric Emergency Patients'Waiting Time Based on Machine Learning Algorithms GUO Linlin, WEI Yongxiang, GUO Linying, Children's Hospital, Capital Institute of Pediatrics, Beijing 100020, China; LI Jiao, ZHENG Si, Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China; CUI Ying, CHEN Yu, Children's Hospital, Capital Institute of Pediatrics, Beijing 100020, China

〔Abstract〕 Based on emergency patients' records of Children's Hospital, Capital Institute of Pediatrics, 19 visit time-related variables are designed to construct 6 different machine learning models to predict the waiting time of patients in pediatric emergency department. The test results show that the prediction models perform well. This study helps hospital managers to dynamically allocate medical resources, relieve crowdedness in emergency departments, and improve patient satisfaction.

〔Keywords〕 pediatric emergency; machine learning; waiting time; prediction

1 引言

儿科急诊患儿候诊时间长是医患关系的主要矛盾点之一。患儿就医往往由 2~3 名成人陪同, 拥挤的环境和焦躁的候诊情绪不仅降低患者就医体验, 也增加急诊科医护人员的职业紧张^[1]。医院所采取的优化就医流程相关措施, 如增加预检分诊护

〔修回日期〕 2022-01-24

〔作者简介〕 郭琳琳, 硕士, 副研究员, 发表论文 13 篇; 通讯作者: 魏永祥, 博士, 院长。

〔基金项目〕 首都儿科研究所所级基金 (项目编号: SK-2021-04)。

士和看诊医生排班等, 根据管理者经验判断实施, 无法科学、精准地预测患者候诊时间, 动态调配医疗资源。患者候诊、就诊时间是衡量急诊科效率、评估当前拥挤状态的最常用指标^[2]。评估患者候诊时间有利于医院管理人员合理分配医疗资源, 使急诊患儿及时得到救治, 避免拥挤聚集情况出现。因此患者候诊时间预测是缓解急诊科拥挤状况的重要方法和手段。当前针对儿科急诊患者候诊时间预测的相关研究较少。本研究以首都儿科研究所附属儿童医院 2021 年 1-6 月共 95 875 条数据为基础, 提出一套基于机器学习的患者候诊时间预测模型的标准构建流程, 包括原始数据提取、异常值检测、候诊时间相关变量设计及提取、描述性统计分析、机器学习模型构建、模型评价及特征重要性分析。

2 研究现状

2.1 基于量表评分的方法

目前评估医院拥挤程度的主流方法。该方法通过预定义一组与医院医疗资源和患者情况相关的变量集合, 使用加权求和的方法计算拥挤程度分级。国家急诊科过度拥挤研究 (National Emergency Department Overcrowding Study, NEDOCS) 评分系统被广泛应用于评估急诊科拥挤程度, 该评分系统包含 23 个指标并通过计算给出拥挤程度评估结果, 最初研究在两个医院中取得 0.49 的决定系数^[3]。严重过度拥挤-过度拥挤-未过度拥挤 (Severely overcrowded-Overcrowded-Not overcrowded Estimation Tool, SONET) 是用于评估急诊科过度拥挤程度的工具, SONET 考虑当前急诊科中不同危重分级、状态的患者人数并加权计算得出拥挤度评分, 结果显示该评分优于广泛使用的 NEDOCS 评分^[4]。然而该方法预测性能受限, 仅能给出当前医院的拥挤程度, 无法计算具体等待时间预测结果。

2.2 基于离散事件模拟的方法

将患者就诊问题转换为排队问题, 使用仿真软件模拟排队行为, 从而进行等待时间预测。Hung G R、Whitehouse S R 和 O'Neill C 等^[5]使用离散时间建

模软件构建患者流动模型用于模拟儿科急诊科的人员配置场景, 结果表明该模型响应人员配置变化并对等待时间预测结果做出相应改变; Connelly L G 和 Bair A E^[6]使用 Extend DES 建模软件模拟急诊科活动模型, 通过输入医疗人员配备水平、设施特征、患者数据、账单记录等信息预测患者平均等待时间, 结果表明 58% 患者的绝对误差小于 3 小时。然而, 该方法较为复杂, 需要较多历史数据验证仿真模型, 同时需要设置较多参数以近似真实情况。

2.3 基于机器学习的方法

根据患者就诊时间相关变量, 采用线性回归、决策树等机器学习算法对其进行建模并预测具体的候诊时间。该方法的核心是构建合适且相关的患者变量。Kuo Y H、Chan N B 和 Leung J M Y 等^[7]使用两组变量集合和 4 种机器学习模型构建急诊科候诊时间预测模型, 其中最优模型在测试集上的决定系数为 0.429; Sun Y、Teow K L 和 Heng B H 等^[8]使用分位数回归算法构建急诊科患者候诊时间预测模型, 在中位候诊时间为 17 分钟的数据集中模型预测绝对误差小于 20 分钟; Pak A、Gannon B 和 Staib A^[9]使用 4 种回归算法和随机森林模型用于急诊科候诊时间预测, 结果表明机器学习算法优于简单的滚动平均方法。何跃、邓唯茹和刘司寰^[10]使用 3 种决策树组合 BP 神经网络构建急诊等待时间预测模型, 但是该方法缺少变量与等待时间相关性的进一步探究。刘翠、王巍和代伟等^[11]使用人工神经网络构建超声科患者排队时间预测模型, 发现患者排队时间的变化规律。通过医院工作流程的精细化管理和医疗资源的合理分配来缩短患者就诊时间, 可以提高患者满意度^[12-13]。结合本院数据和有关研究的样本^[9]发现, 来源于不同地区的不同人群的急诊患者候诊时间数据会存在一定差异, 因此需要选取合适的机器学习模型来进行研究。

3 资料与方法

3.1 数据来源

本院 2021 年 1-6 月共 95 875 条急诊患者就诊

数据，包括患者性别、年龄、医生工资号、就诊分区、取号时间、首诊时间、首诊名称、首次医嘱时间、末次医嘱时间、是否有辅助检查（检验、检查等）及是否开药。

3.2 数据预处理

对原始数据进行数据清洗、加工处理，以得到适合候诊时间预测建模的数据。本研究主要采用以下 3 条数据清洗规则：去除缺失急诊分区信息的样本数据；去除候诊时间小于 0（首诊时间到取号时间小于 0 分钟）或者年龄大于 18 周岁的错误样本数据；去除总就诊时间异常（末次医嘱时间到取号时间大于 2 天）和候诊时间异常（首诊时间到取号时间大于 480 分钟）的样本数据。经过数据清洗后，最终得到 90 573 条样本数据用于后续分析，见图 1。

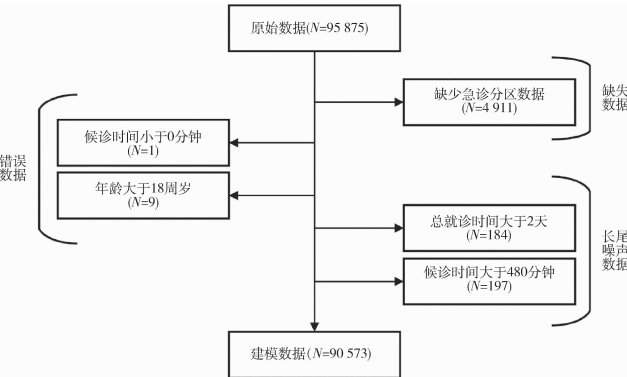


图 1 数据清洗流程

3.3 描述性分析和特征提取

对预处理后的样本数据进行描述性统计分析，以发现不同临床亚群患者相关特征，辅助指导构建等待时间预测建模相关变量设计。此外通过绘制就诊人数、患者平均候诊时间等随时间变化的趋势图，帮助医院管理人员了解就诊人数和急诊科拥挤程度在不同时间段的变化模式并以此实施相应决策。对于特征提取方面，根据先前研究^[7,9]，本研究设计并提取 19 个变量用于候诊时间预测模型构建，见表 1。

表 1 变量描述

变量名称	变量解释
wait. t	等待时间（首诊时间到取号时间）
arrivals. green	患者取号时，前 1 小时绿区到达人数
arrivals. other	患者取号时，前 1 小时红区和黄区到达人数
left. green	患者取号时，前 1 小时绿区离院人数
left. other	患者取号时，前 1 小时红区和黄区离院人数
treat. green	患者取号时，绿区正在就诊人数
treat. other	患者取号时，红区和黄区正在就诊人数
wait. green	患者取号时，绿区正在等待人数
wait. other	患者取号时，红区和黄区正在等待人数
4h. green	患者取号时，前 4 小时绿区就诊完毕人数
4h. other	患者取号时，前 4 小时红区和黄区就诊完毕人数
t. avg4h. green	患者取号时，前 4 小时绿区就诊完毕患者的平均等待时间
t. avg4h. other	患者取号时，前 4 小时红区和黄区就诊完毕患者的平均等待时间
doctor	当前值班医生数
month	患者取号时间段（月份）
week	患者取号时间段（星期）
holiday	患者取号时间段（是否节假日）
triage	患者取号时的分区
rain	患者取号当天是否存在极端天气（下雨、沙尘等）

3.4 基于机器学习的候诊时间预测模型构建

使用线性回归（Linear Regression, LR），回归模型（Least Absolute Shrinkage and Selection Operator, LASSO），K 最近邻（K - Nearest Neighbor, KNN），随机森林（Random Forest, RF），极端梯度森林^[14]（eXtreme Gradient Boosting, XGBoost）及 LightGBM^[15]共 6 种机器学习算法构建候诊时间预测模型。LR 是一种利用最小二乘法将多个自变量逼近拟合因变量的回归分析算法，可以简易地表示为 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon$ 。LASSO 是一种在 LR 的基础上进行改进的方法，其在最小二乘法中引入 L1 惩罚项以压缩数据空间维度，从而实现参数估计和变量选择。KNN 是一种在样本空

间上寻找 K 个最近邻样本作为待预测样本的估计，并计算近邻样本标签均值作为预测值的算法。RF 是一种包含多个决策树的集成机器学习算法。其使用 Bootstrap 方法从训练集中有放回地抽样，并对其每个样本随机选择一定数量特征构建多个决策树，最后在预测阶段将所有决策树预测结果的平均值作为最终预测结果。XGBoost 是一种基于梯度提升决策树的增强算法。其在训练时使用贪婪策略，在每次迭代时都学习一个决策树用于预测上一个决策树预测值与真实值的残差；同时，其将二阶泰勒公式展开引入损失函数以提高精确度，并利用正则项简化模型。LightGBM 是一种基于梯度提升决策树的增强算法。其采用单边梯度采样、互斥特征捆绑、受限决策树生长策略等方法降低模型复杂度，加速模型训练速度的同时使模型精确度保持稳定。

3.5 实验参数设置

首先对 6 种机器学习模型进行超参数网格搜索，以确定每种机器学习模型的最优超参数，见表 2。采用 10 折交叉验证用于评估模型表现，设置 10 个随机种子以重复实验，给出模型表现的平均值。对于模型表现结果，采用决定系数 R^2 ，平均绝对误差（Mean Absolute Error, MAE），平均绝对百分误差（Mean Absolute Percentage Error, MAPE），均方根误差（Root Mean Square Error, RMSE）作为评价指标，计算方法为：

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{1}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \tag{2}$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \tag{4}$$

其中 y_i 、 $\hat{y}_i$ 、 $\bar{y}$ 分别为样本真实值、样本预测值、真实值的均值。

表 2 机器学习模型超参数

模型	模型超参数
LR	Fit_ intercept: [True, False]; Normalize: [True, False]
Lasso	Fit_ intercept: [True, False]; Normalize: [True, False]; Alpha: [0.001, 0.01, 0.1, 1, 10, 100]
KNN	N_ neighbors: [3, 4, 5, 6, 8]; Leaf_ size: [10, 15, 20, 25, 30]
RF	N_ estimators: [100, 200, 250, 300, 350, 400]; Max_ depth: [10, 15, 17, 20, 22, 25]
XGB	N_ estimators: [80, 100, 150, 200, 250, 300, 350, 400, 450]; Learning_ rate: [0.01, 0.05, 0.07, 0.08, 0.09, 0.1]; Max_ depth: [5, 8, 10, 12, 14, 16, 18, 20]

4 结果

4.1 儿科急诊拥挤程度描述性统计

首先对 90 573 条样本数据的关键信息进行描述性统计。可知绿区占总就诊人数最多，而红区、黄区患者较少。由于就诊分区为红区的患者病情最危重，在实际情况中均为立即就诊，因此不需要对其进行候诊时间预测，但由于其在一定程度上与黄区、绿区患者共同使用医疗资源，因此仍加入红区患者参与计算。此外夜班（20 时 ~ 次日 8 时）中的平均就诊人数略多于白班（8 时 ~ 20 时），而夜班医生数略少于白班医生数。门诊小夜班结束后，病情较轻患者只能在急诊科就医，也会对急诊医疗资源产生一定程度消耗，见表 3。

表 3 患者就诊数据描述性统计

统计项		总计
样本数	红区	2 460 (2.71%)
	黄区	2 742 (3.03%)
	绿区	85 373 (94.26%)
	平均年龄	4.08 ± 3.23
	男性比例	51 433 (56.79%)
平均等待时间 (SD)	红区	18.84 ± 37.31
	黄区	20.28 ± 48.81
	绿区	70.58 ± 84.51
平均等待人数 (SD)	白班	32.05 ± 30.66
	夜班	37.17 ± 33.71
平均值班医生数 (SD)	白班	15.53 ± 2.42
	夜班	12.97 ± 2.56

4.2 儿科急诊候诊时间趋势分析

根据 2021 年 1-6 月就诊数据进行时间趋势分析可知，急诊科拥挤程度（患者候诊时间）与就诊人数高度相关，在 2021 年 1-3 月，由于防疫政策、

节假日、季节等因素影响，急诊科总体就诊人数较少，同时患者平均等待时间短（低于 25 分钟）；之后随着防疫政策调整、人员流动较年初更加自由、学校开学、逐步入夏等原因，急诊科就诊人数上升，同时患者平均候诊时间也大幅上升，最高近 3 小时。由此可知不同时间段急诊科拥挤程度变化较大。为了最有效缓解该现象，需要根据患者候诊时间灵活实施加速就诊措施。因此患者候诊时间预测模型是为执行各种措施提供时间预估的重要手段。

4.3 就诊候诊时间预测模型表现

4.3.1 模型构建 使用 LR、LASSO、KNN、RF、XGBoost、LightGBM 构建基于机器学习的急诊患者候诊时间预测模型。所有模型的整体预测能力均较好，其中 LightGBM 取得最优的 R^2 、MAE 及 RMSE，相对于作为基线水平的 LR 模型，LightGBM 在 3 个指标上的提升分别为 8.28%、9.54% 和 4.28%，模型表现优越，可以对不同状态下的患者提供较为准确的候诊时间预测，见表 4。

表 4 机器学习模型表现结果（分钟）

模型	模型表现			
	R^2	MAE	MAPE	RMSE
LR	0.483 ± 0.000	40.434 ± 0.001	364.604 ± 0.048	62.092 ± 0.003
LASSO	0.483 ± 0.000	40.397 ± 0.001	363.542 ± 0.047	62.096 ± 0.003
KNN	0.487 ± 0.001	37.731 ± 0.026	335.205 ± 0.466	61.871 ± 0.046
RF	0.519 ± 0.000	36.802 ± 0.010	326.792 ± 0.324	59.892 ± 0.020
XGBoost	0.520 ± 0.000	37.054 ± 0.006	322.439 ± 0.335	59.846 ± 0.012
LightGBM	0.523 ± 0.000	36.576 ± 0.014	323.739 ± 0.396	59.642 ± 0.027

4.3.2 6 种机器学习模型在不同时间段的等待时间拟合能力分析 绘制 24 小时段绿区、黄区患者的平均候诊时间趋势图，观察结果可知急诊科患者平均候诊时间存在两个高峰，分别为 8~12 时、20~23 时，该结果符合本院实际诊疗情况。此外 6 个机器学习模型均能较好拟合并预测该变化趋势，平

均候诊时间的 MAE 分别为 13.071、13.082、13.180、13.319、14.448、13.376。证明使用机器学习算法构建患者候诊时间预测模型的稳健性和准确性较好；同时 LightGBM 表现出最优的拟合效果，见图 2。

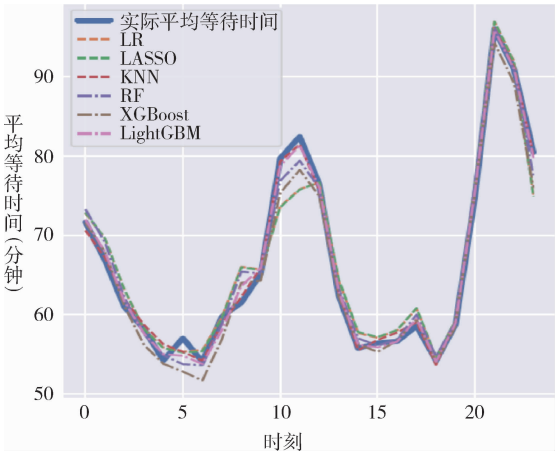


图2 患者平均候诊时间预测（24 小时）

4.3.3 不同分区患者平均等待时间预测分析 由于不同分区患者代表不同病情危重程度的临床人群，因此候诊时间预测模型需要区分不同分区患者以促进个性化诊疗实施。通过绘制不同分区患者平均候诊时间预测结果的时间趋势可知，虽然 6 种机器学习模型均能准确预测绿区患者的平均候诊时间变化，但 LR、LASSO、KNN 模型对于黄区患者几乎无预测能力，而 3 种决策树集成模型（RF、XGBoost、LightGBM）在黄区患者中仍有较为准确的预测能力。该结果表明在患者候诊时间预测中，复杂度较高的集成模型能够学习到罕见临床亚群信息，从而提供泛化性更好的预测，见图 3。

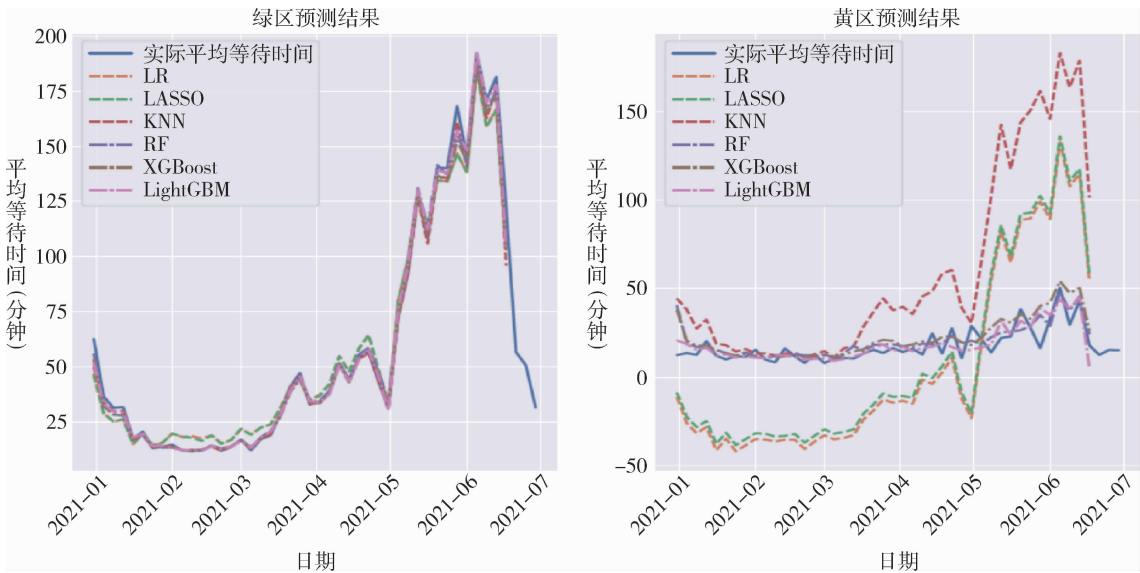


图3 不同分区患者平均等待时间预测时间曲线

4.4 特征重要性分析

使用 6 种机器学习模型中表现最优的 LightGBM 模型输出其特征重要性分数。观察结果可知，本研究设计的多个变量（wait. green、t. avg4h. green、arrivals. green 等）在 LightGBM 模型中重要程度较高，而原始数据中的部分变量（doctor、triage、holiday 等）对结果影响较小。以上结果反映了本研究设计的变量集合对患者等待时间预测模型构建的重要性；同时该结果也证明诊疗活动的历史信息能够用于预测当前及未来医院拥挤程度，从而辅助医院管理者决策，见图 4。

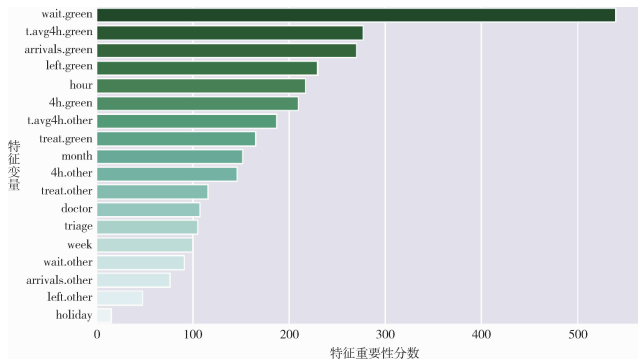


图4 特征重要性分数排序

5 结语

提高患者就诊效率和满意度是在儿科急诊诊疗活动中执行医疗资源管理、分配的重要目标。其中针对儿科急诊患者候诊时间进行预测是科学安排医疗资源以缓解急诊科拥挤程度的前提。本研究基于大规模患者就诊数据共构建 6 种不同机器学习模型用于儿科急诊就诊患者的候诊时间预测, 最佳模型的决定系数 (R -squared, R^2) 为 0.523, 预测模型表现良好, 能够较为精准地预测急诊科患者候诊时间, 同时精确拟合不同时间段下患者平均候诊时间, 具备较高稳健性和泛化性。综上, 该模型有助于医院管理者提前采取应对措施以合理分配医疗资源, 缓解急诊科拥挤程度, 改善医疗环境, 提高患者满意度。本研究拟于下一步将该模型应用至本院急诊科并探讨其在其他诊疗科室的应用潜力, 从而在真实场景下帮助医院管理者更高效、合理地实施资源动态调配。研究存在一定局限性。研究仅限于 2021 年 6 个月的样本数据, 无法反映年度季节性趋势, 如月份, 对于候诊时间的影响。在未来的研究中将采用更加完整的年度样本数据以及考虑增加疾病诊断和治疗等临床因素, 开展急诊候诊时间更深入的比较分析, 使结果更加全面。

参考文献

- 1 高妍, 赵俊琴, 赵春香, 等. 急诊科医护人员职业紧张的研究进展 [J]. 职业与健康, 2021, 37 (18): 2568-2570, 2576.
- 2 Aldridge E S, Rogers I R, Bailey P M, et al. Emergency Department 'undercrowding' is Associated with Decreased Waiting Times [J]. Emerg Med Australas, 2016, 28 (3): 268-272.
- 3 Weiss S J, Derlet R, Arndahl J, et al. Estimating the Degree of Emergency Department Overcrowding in Academic Medical Centers: Results of the National ED Overcrowding Study (NEDOCS) [J]. Acad Emerg Med, 2004, 11 (1):

- 38-50.
- 4 Wang H, Robinson R D, Garrett J S, et al. Use of the SO-NET Score to Evaluate High Volume Emergency Department Overcrowding: A Prospective Derivation and Validation Study [J]. Emerg Med Int, 2015 (2015): 401757.
- 5 Hung G R, Whitehouse S R, O'Neill C, et al. Computer Modeling of Patient Flow in a Pediatric Emergency Department Using Discrete Event Simulation [J]. Pediatr Emerg Care, 2007, 23 (1): 5-10.
- 6 Connelly L G, Bair A E. Discrete Event Simulation of Emergency Department Activity: a Platform for System-level Operations Research [J]. Acad Emerg Med, 2004, 11 (11): 1177-1185.
- 7 Kuo Y H, Chan N B, Leung J M Y, et al. An Integrated Approach of Machine Learning and Systems Thinking for Waiting Time Prediction in an Emergency Department [J]. Int J Med Inform, 2020 (139): 104143.
- 8 Sun Y, Teow K L, Heng B H, et al. Real-time Prediction of Waiting Time in the Emergency Department, Using Quantile Regression [J]. Ann Emerg Med, 2012, 60 (3): 299-308.
- 9 Pak A, Gannon B, Staib A. Predicting Waiting Time to Treatment for Emergency Department Patients [J]. Int J Med Inform, 2021 (145): 104303.
- 10 何跃, 邓唯茹, 刘司寰. 基于组合决策树的急诊等待时间预测 [J]. 统计与决策, 2016 (6): 72-74.
- 11 刘翠, 王巍, 代伟, 等. 武汉市某三甲医院超声科患者就诊情况分析 with 排队时间预测 [J]. 医学与社会, 2020, 33 (2): 45-49.
- 12 乔林楠, 陈蕾, 方碧兰. 精益管理在降低心脏电生理检查患者候诊时间中的应用 [J]. 中医药管理杂志, 2021, 29 (16): 251-252.
- 13 张丽, 孙宾, 孟永前. 超声科患者就诊情况分析 with 排队时间预测 [J]. 中医药管理杂志, 2021, 29 (14): 223-225.
- 14 Chen T, He T, Benesty M, et al. Xgboost: Extreme Gradient Boosting [J]. R Package Version 0.4-2, 2015, 1 (4): 1-4.
- 15 Ke G, Meng Q, Finley T, et al. Lightgbm: A Highly Efficient Gradient Boosting Decision Tree [J]. Advances in Neural Information Processing Systems, 2017 (30): 3146-3154.