

· 临床研究 · doi:10.3969/j.issn.1671-8348.2024.22.003

网络首发 [https://link.cnki.net/urlid/50.1097.R.20240903.1202.019\(2024-09-03\)](https://link.cnki.net/urlid/50.1097.R.20240903.1202.019(2024-09-03))

ARMA 模型和 LSTM 深度神经网络对新疆伽师县肺结核发病趋势预测效果的比较^{*}

穆妮热·克日木¹, 买吾拉江·依马木², 美合日班·买买提¹, 张丽萍³, 郑彦玲^{3△}

(1. 新疆医科大学公共卫生学院, 乌鲁木齐 830017; 2. 喀什地区疾病预防控制中心, 新疆喀什 844000;

3. 新疆医科大学医学工程技术学院, 乌鲁木齐 830017)

[摘要] **目的** 运用自回归移动平均(ARMA)模型和长期短期记忆(LSTM)深度神经网络对新疆伽师县肺结核的发病趋势进行预测。**方法** 收集该地 2014 年 1 月至 2023 年 6 月法定传染病报告肺结核数据构建数据集, 其中 2014 年 1 月至 2021 年 12 月的肺结核发病数据用于模型构建, 2022 年 1 月至 2023 年 6 月的数据用于模型验证。利用 Eviews7.2 和 MATLAB2023a 软件分别建立 ARMA 模型和 LSTM 神经网络, 预测 2022—2023 年的月肺结核发病数。**结果** 最优的 ARMA 模型和 LSTM 神经网络验证 2014 年 1 月至 2023 年 6 月发病数的均方根误差(RMSE)分别为 26.494 和 12.713, 提示 LSTM 神经网络的拟合效果优于 ARMA 模型。采用 LSTM 神经网络预测结果与实际发病情况基本一致。**结论** LSTM 神经网络对新疆伽师县发病趋势的拟合和预测效果较好, 能够为该地肺结核发病数的预测提供理论参考。

[关键词] 肺结核; ARMA 模型; LSTM 模型; 预测

[中图分类号] R521

[文献标识码] A

[文章编号] 1671-8348(2024)22-3375-05

Comparison between ARMA model and LSTM deep neural network in predictive effect on onset trend of pulmonary tuberculosis in Jiashi County of Xinjiang^{*}

MUNIRE Kerimu¹, MAIWULAJIANG Yimamu², MEIHERIBAN Maimaiti¹,

ZHANG Liping³, ZHENG Yanling^{3△}

(1. School of Public Health, Xinjiang Medical University, Urumqi, Xinjiang 830017, China; 2. Kashi Prefecture Disease Control and Prevention Center, Kashi, Xingjiang 844000, China; 3. College of Medical Engineering Technology, Xinjiang Medical University, Urumqi, Xinjiang 830017, China)

[Abstract] **Objective** To use the auto-regressive moving average (ARMA) model and long short term memory (LSTM) depth neural network to predict the incidence trend of pulmonary tuberculosis in Jiashi County. **Methods** The legal infectious disease report data in this area from January 2014 to June 2023 were collected to construct the data set, in which the onset data of pulmonary tuberculosis from January 2014 to December 2021 were used to the model construction and the data from January 2022 to June 2023 were used to the model verification. The Eviews7.2 and MATLAB2023a softwares were used to construct the ARMA mode and LSTM neural network. The monthly onset number of pulmonary tuberculosis from 2022 to 2023 was predicted. **Results** The root-mean-square error (RMSE) of the optimal ARMA model and LSTM neural network verification from January 2014 to June 2023 was 26.494 and 12.713 respectively, suggesting that the fitting effect of LSTM neural network was better than that of ARMA model. The predictive results by adopting the LSTM neural network was basically consistent with the actual onset situation. **Conclusion** The LSTM neural network has good fitting and predicting effect for the onset trend in Jiashi County, which could provide the theoretical reference for predicting the onset number of pulmonary tuberculosis in the future in this area.

[Key words] tuberculosis; ARMA model; LSTM model; forecast

结核病是一种严重的慢性传染病^[1], 可侵及多种器官, 其中肺结核占 90% 以上^[2]。结核病是全球面临

^{*} 基金项目: 国家自然科学基金项目(72064036, 72174175, 72163033)。[△] 通信作者, E-mail: zhengyl_math@sina.cn。

的公共卫生问题,《2022 全球结核病报告》指出,2021 年我国报告肺结核病例 639 548 例,死亡病例 1 763 例,报告发病率为 45.37/10 万,报告死亡率为 0.13/10 万^[3],在全球结核病高负担国家中排名第 3 位^[4]。新疆喀什地区是肺结核的高发地区之一,其结核病发病率常年高于全国平均水平数倍,2011—2020 年的年均报告发病率达到了 451.29/10 万^[5],防控形势严峻。

预测传染病的发病趋势有助于及时采取有效措施预防和控制疾病流行。目前传染病发病预测模型主要有线性模型、非线性模型及混合模型^[6]。自回归滑动平均模型(auto-regressive moving average model, ARMA)作为经典的预测模型,已被应用于传染病及其他学科^[7]。近年来,深度机器学习逐渐被应用在时间序列的分析预测方面,其中循环神经网络(recurrent neural network, RNN)表现较好;而基于 RNN 的长期短期记忆(long short term memory, LSTM)网络解决了 RNN 在分析预测时序数据中出现梯度消失、梯度爆炸及长期记忆能力不足的问题,可更为有效地处理分析长时间序列数据^[8]。本研究对 2014—2023 年新疆喀什地区伽师县报告的肺结核发病数据进行流行病学分析,并根据既往月发病数建立预测发病数的 ARMA 和 LSTM 模型,对模型进行评价,以确定更适合该地肺结核发病数预测的模型,为后续提供科学理论参考。

1 资料与方法

1.1 一般资料

数据来自新疆维吾尔自治区喀什地区疾病预防控制中心,收集 2014 年 1 月至 2023 年 6 月现住址为喀什地区伽师县的肺结核患者信息,包括性别、年龄、住址等。其中 2014 年 1 月至 2021 年 12 月的肺结核发病数据用于模型构建,2022 年 1 月至 2023 年 6 月的数据用于模型验证。

1.2 方法

1.2.1 ARMA 模型

ARMA 模型的核心思想是将预测目标随时间变化形成的数据序列视作随机过程,并在此基础上利用时间序列的自相关分析,通过构建数学模型来捕捉序列中的自相关性特征,通过这种方式,模型能够识别出最优的参数配置,并利用这些参数来预测未来的数据点。ARMA 模型是由 AR 自回归过程和 MA 滑动平均过程组成的自回归滑动平均模型,是时间序列分析模型的一种,其基本思想是通过揭示历史时间序列的运行规律,从而对未来的事物发展进行预测。在 ARMA(p, q)模型中, p 为自回归阶数, q 为滑动平均

阶数^[9]。建立 ARMA 模型主要包括以下 4 步骤:(1)平稳性检验:首先通过绘制时序图、自相关图和偏自相关图或进行单位根检验(augmented dickey-fuller test, ADF)来判断数据的稳定性。(2)模型的识别:包括建立模型类型和确定相应阶的值,所识别的工具为自相关函数(ACF)和偏自相关函数(PACF),确定 p 和 q 。根据平稳序列的 ACF 和 PACF 图,固定参数(p, q),建立合理的模型。当 ACF 图截断而 PACF 图拖尾时, q 值对应 ACF 图中被截断的值。当 ACF 图拖尾而 PACF 图截断时, p 值对应 PACF 图中被截断的值^[10],并按照从高到低的顺序选择模型并最终确定出 q, p 的最佳值^[11]。(3)模型参数估计:赤池信息准则(Akaike information criterion, AIC)、施瓦茨准则(Schwarz critrion, SC)值最小的模型为最优模型;同时,稳定 R^2 越大,模型的拟合效果越好。(4)模型的诊断一般采用 Box-Ljung Q 统计量,当这个统计量无统计学意义($P>0.05$),说明残差是白噪声,所选的模型可靠。根据 AIC 来衡量所选模型与数据适合度,通常模型越精细、适合度越好,即 AIC 越小、模型越适合预测。

1.2.2 LSTM 模型

LSTM 是基于 RNN 的一种形式,与 RNN 不同之处在于 LSTM 模型增加了门,解决了 RNN 在层数较多时出现梯度弥散的问题,LSTM 模型对处理肺结核发病数间序列时有较好的效果^[12]。一个典型的 LSTM 单元包括输入门、遗忘门和输出门三种门^[13],见图 1, LSTM 通过第一步遗忘门 f_t 来决定前一时刻的单元状态 c_{t-1} 流动到 t 时刻后丢弃什么信息,当遗忘门的值为 1 时,它保留该信息;当值为 0 或接近 0 意味着去掉了所有信息。另一个是输入门,输入门 i_t 它选择性确定当前时刻的网络的输入 x_t 被存放在单元状态中。最终 LSTM 使用输出门 o_t 来确定单元状态 o_t 哪些部分输出到 LSTM 的当前隐含状态 h_t 中,最终得到 h_t 作为下一步时间输入的一部分^[14]。

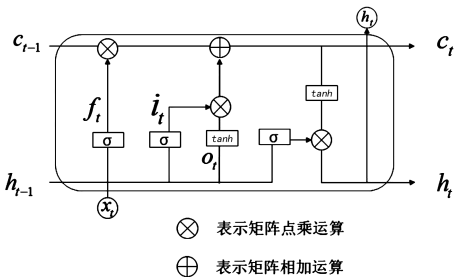


图 1 LSTM 神经网络的单元结构

2 结 果

2.1 流行病学特征

2.1.1 性别、年龄分布特征

男 4 556 例,占比 50.51%,而女 4 463 例,占比 49.49%,男女比例为 1.021 : 1。肺结核的发病率随着年龄的增长而增加,特别是>60~70 岁的上升趋势最为明显。>70~80 岁年龄组肺结核的发病率最高,达到 29.10%。在>20~30 岁、>30~40 岁、>40~50 岁、>70~80 岁及>80 岁年龄组中,男多于女,见表 1。

表 1 伽师县肺结核患者年龄、性别分布情况[n(%)]			
项目	合计	男	女
合计	9 019(100.00)	4 556(50.51)	4 463(49.49)
≤20 岁	168(1.86)	68(1.50)	100(2.24)
>20~30 岁	618(6.85)	337(7.40)	281(6.30)
>30~40 岁	635(7.04)	395(8.67)	240(5.38)
>40~50 岁	527(5.84)	292(6.41)	235(5.27)
>50~60 岁	1 018(11.29)	424(9.31)	594(13.31)
>60~70 岁	2 422(26.85)	1 067(23.42)	1 355(30.36)
>70~80 岁	2 623(29.08)	1 371(30.09)	1 252(28.05)
>80 岁	1 008(11.18)	602(13.21)	406(9.10)

2.1.2 地区分布特征

伽师县 13 个乡镇中克孜勒博依乡、夏阿瓦提乡、江巴孜乡和英买里乡等 4 个乡镇每年肺结核发病人数占伽师县的大部分,见图 2。

2.2 ARMA 模型的建立

对 2014 年 1 月至 2023 年 6 月伽师县肺结核逐月发病数绘制时序图,呈现为先上升后下降、然后小幅波动的趋势,总体上是平稳的时间序列,见图 3。2014 年 1 月至 2023 年 6 月伽师县总发病 9 019 例,月发病数波动较大但没有明显规律。

ACF 和 PACF 是识别纯 AR(p)或 MA(q)模型的有效工具,ACF 和 PACF 均是拖尾的,见图 4,因此判断该数据适合做 ARMA(p, q)模型。考虑到 ACF

和 PACF 确定的 ARMA 模型阶次具有一定的主观性,参数 p 、 q 分别取 0、1、2 得到几个不同的 ARMA 模型并进行进一步分析。经 Eviews7.2 测算,模型 ARMA(1,1),ARMA(1,2),ARMA(2,1),ARMA(2,2)均通过了检测,系数均为显著。在此基础上对模型的 AIC 及 SC 值进行比较,结果见表 2,ARMA(1,1)模型的 AIC 和 SC 值分别取 9.427、9.475 时最小,表明 ARMA(1,1)为最优模型。用 ARMA(1,1)模型拟合和预测伽师县肺结核发病数比较接近实际情况,表明模型预测效果较好,见图 5。

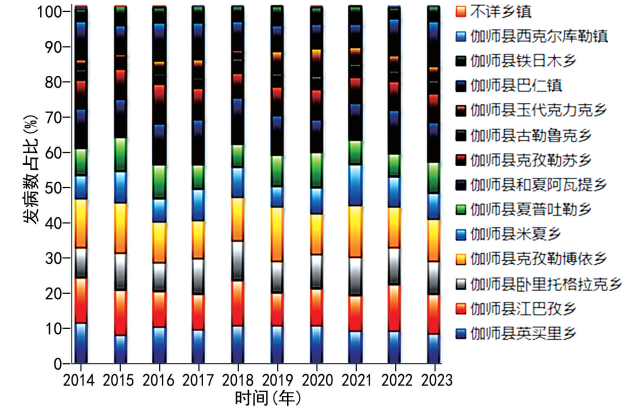


图 2 伽师县各个乡镇 2014—2023 年肺结核发病例数占比图

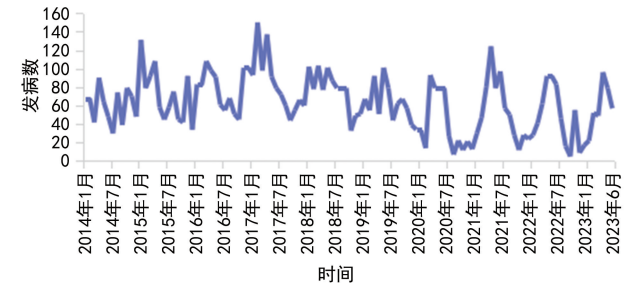


图 3 2014 年 1 月至 2023 年 6 月伽师县肺结核逐月发病数时序图

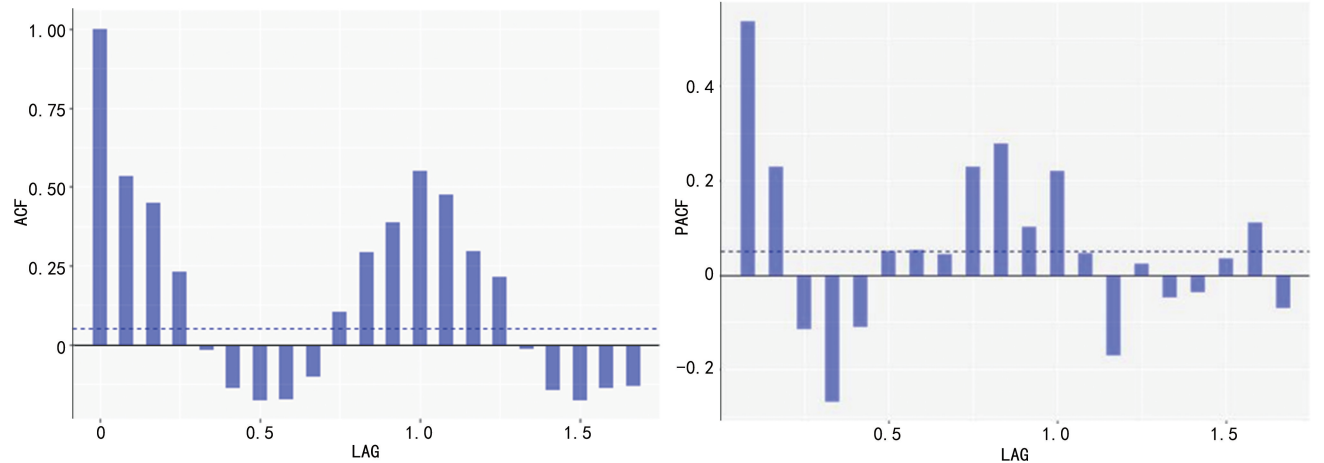


图 4 ACF 和 PACF 函数图

表 2 ARMA(1,1)模型参数估计和拟合优度统计量结果

变量	AIC	SC
ARMA(1,1)	9.427	9.475
ARMA(1,2)	9.504	9.552
ARMA(2,1)	9.509	9.557
ARMA(2,2)	9.728	9.776

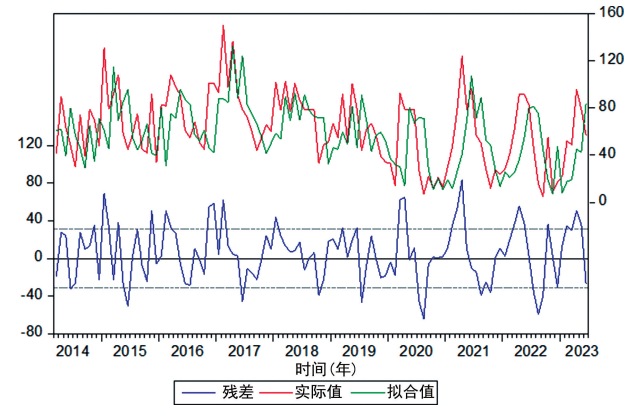


图 5 ARMA(1,1)模型对 2014 年 1 月至 2023 年 6 月伽师县肺结核逐月发病数拟合及预测效果

2.3 LSTM 模型的建立

采用 LSTM 模型,使用 2014 年 1 月至 2021 年 12 月伽师县肺结核月发病数数据构建训练集,2022 年 1 月至 2023 年 6 月的月发病数据用作验证及预测。LSTM 模型月度拟合值与伽师县肺结核的实际发病数非常接近,说明 LSTM 模型的预测效果较优,见图 6。

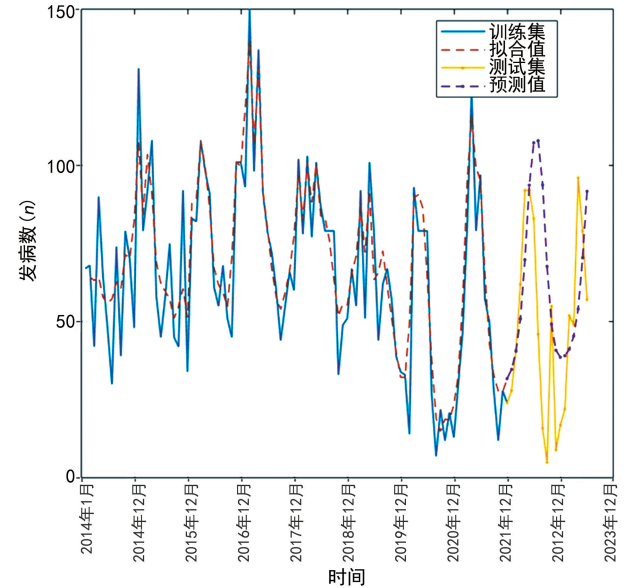


图 6 LSTM 模型对 2014 年 1 月至 2023 年 6 月伽师县肺结核逐月发病数的拟合及预测效果

2.4 模型拟合效果比较

通过均方根误差(Root-mean-square error, RMSE)定量分析两种模型的拟合和预测效果,对模型

精确度做出评价。ARMA 模型和 LSTM 神经网络模型拟合及预测的发病数与实际发病数之间的 RMSE、 R^2 值可以看出 LSTM 神经网络模型的拟合和预测效果明显优于 ARMA(1,1)模型,说明 LSTM 神经网络模型在伽师县 2014—2023 年肺结核发病数据预测应用中表现良好,见表 3。

表 3 两种模型预测评价

项目	预测评价	
	RMSE	R^2
ARMA(1,1)模型	26.494	0.216
LSTM 模型	12.713	0.813

3 讨 论

预测传染病发病能够帮助理解并掌握传染病的传播和发展规律,对有效预防和控制传染病至关重要。目前有多种预测传染病的模型,包括 ARIMA 模型、GM(1,1)模型、灰色模型、Markov 模型、人工神经网络模型和通径分析模型等^[15-16]。使用 ARIMA 模型对肺结核进行短期预测是有效的,这对于及时发现慢性传染病的爆发或流行非常关键,有助于提前进行针对性地健康教育和干预^[17]。LSTM 神经网络因其对处理时间序列数据中的多重相关性和非线性问题的适用性,相较于传统的统计分析模型,显示出更大的优势^[8]。2011—2020 年,新疆地区肺结核患者的登记率整体呈上升趋势,平均每年发病率达到 159.46/10 万,是全国平均水平的 2.77 倍^[18]。喀什地区是新疆结核病疫情高发地区之一^[19],肺结核报告发病率高于其他地(州)。2011—2020 年喀什地区肺结核报告发病率呈先上升后下降趋势,最低在 2020 年^[5]。ARMA 模型和 LSTM 模型对预测伽师县肺结核月发病数是可行的,但需要强调的是,与前几年相比,2022 年发病数明显下降,这可能与新型冠状病毒感染疫情带来的影响有关:来自偏远地区的居民减少了就医频率或延误就医,因此没有得到及时有效的诊断,影响了新病例的报告^[20-21]。本研究存在的局限性在于肺结核预测模型众多,如 BP 神经网络、Holt-Winters 指数平滑法、Prophet 模型等模型,有待进一步研究。

综上所述,本研究分析了伽师县 2014—2023 年肺结核的分布和流行趋势,分别用 ARMA 模型和 LSTM 神经网络构建伽师县肺结核病例数预测模型,并预测 2022 年 1 月至 2023 年 6 月的发病趋势。从模型的预测结果来看,ARMA(1,1)模型预测结果 RMSE 为 26.494,LSTM 模型预测结果的 RMSE 为 12.713。LSTM 模型的预测精度要高于 ARMA(1,1)模型,说明 LSTM 模型预测伽师县肺结核发病数的

效果优于 ARMA(1,1)模型,研究结果对伽师县肺结核疫情的有效防控具有重要的参考价值。

参考文献

[1] 谢加伟,熊华利,蒋惠,等. 2011—2021 年重庆市荣昌区肺结核空间流行特征分析[J]. 实用预防医学,2023,30(4):424-428.

[2] KANG W,DU J,YANG S,et al. The prevalence and risks of major comorbidities among inpatients with pulmonary tuberculosis in China from a gender and age perspective;a large-scale multicenter observational study[J]. Eur J Clin Microbiol Infect Dis,2021,40(4):787-800.

[3] BAGCCHI S. WHO's Global Tuberculosis Report 2022[J]. Lancet Microbe, 2023,4(1):e20.

[4] 梁晨,于佳佳,唐神结. 世界卫生组织《全球结核病报告 2022》解读[J]. 诊断学理论与实践,2023,22(1):21-30.

[5] 地尔木拉提·吐孙,麦维兰江·阿不力米提,刘振江,等. 2011—2020 年新疆喀什地区肺结核流行特征分析[J]. 热带病与寄生虫学,2023,21(2):93-97,101.

[6] 赖晓莹,钱俊. ARIMA-LSTM-XGBoost 加权组合模型在肺结核发病趋势预测的研究[J]. 现代预防医学,2021,48(1):5-9.

[7] 田庆,刘永鹏,张晶晶,等. ARIMA 乘积季节模型在山东省肺结核发病预测中的应用[J]. 山东大学学报(医学版),2021,59(7):112-118.

[8] 高秋菊,周宇畅,赵树青,等. ARIMA 乘积季节模型和 LSTM 深度神经网络对石家庄市手足口病疫情预测效果的比较[J]. 中华疾病控制杂志,2020,24(1):73-78.

[9] 王莹. 基于 ARMA 模型的股票价格的分析及预测[J]. 生产力研究,2021(9):124-127.

[10] ZHAI M,LI W,TIE P,et al. Research on the predictive effect of a combined model of ARIMA and neural networks on human brucellosis in Shanxi Province,China:a time series predictive analysis[J]. BMC Infect Dis,2021,21(1):1-12.

[11] 何伟,王奇,喻文杰,等. 基于 ARMA 和 GM(1,1)模型预测四川省新发现棘球蚴病患者检出率

变化趋势[J]. 中国寄生虫学与寄生虫病杂志,2021,39(2):185-190.

[12] 贺箫楠,宋晓辉,朱鑫. 基于 ARIMA 模型的洛阳市手足口病发病率预测[J]. 现代预防医学,2019,46(3):403-406.

[13] 王淑平,杜敏,罗建伟,等. GM(1,1)模型与 LSTM 神经网络在肺结核发病数预测中的应用[J]. 公共卫生与预防医学,2019,30(5):11-14.

[14] 倪茹玉,胡婉,张恒川,等. ARIMA 乘积季节模型与 LSTM 神经网络模型对我国麻疹发病数预测效果的比较[J]. 现代预防医学,2023,50(1):177-182.

[15] 张娟娟,王云霞,胡方祥,等. 自回归移动平均模型在深圳市宝安区结核病疫情预测中的应用[J/CD]. 新发传染病电子杂志,2022,7(1):47-51.

[16] 黄国宝,黎衍云,吴菲,等. ARIMA 模型和 ARIMA-SVM 模型对上海市 2 型糖尿病患者肺结核发病的预测效果[J]. 复旦学报(医学版),2020,47(6):899-905.

[17] 周扬,梁士杰. ARIMA 乘积季节模型在郑州市肺结核月发病趋势预测中的应用[J]. 中国卫生统计,2021,38(4):554-555.

[18] 王明哲,王乐,祖丽呼玛尔·艾尔肯,等. 2011—2020 年新疆普通肺结核患者流行病学特征分析[J]. 预防医学情报杂志,2022,38(10):1334-1339,1345.

[19] 麦维兰江·阿不力米提,地尔木拉提·吐孙,克尤木·吾布力喀斯木,等. 新疆喀什地区初治肺结核患者成功治疗后复发影响因素分析[J]. 中国防痨杂志,2022,44(11):1148-1153.

[20] WANG X,HE W,LEI J,et al. Impact of COVID-19 pandemic on pre-treatment delays,detection,and clinical characteristics of tuberculosis patients in Ningxia Hui Autonomous Region, China [J]. Front Public Health, 2021, 9:644536.

[21] PAN A,LIU L,WANG C,et al. Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan,China[J]. JAMA,2020,323(19):1915-1923.