疾病数据分析与建模

DOI：10.13406/j.cnki.cyxb.003384

基于ARIMA模型的重庆市流行性感冒预测研究

邹小江[1]，赵　寒[2]，王祈茵[1]，叶孟良[1]

（1. 重庆医科大学公共卫生学院，重庆　400016；2. 重庆市疾病预防控制中心，重庆　400000）

【摘　要】目的：分析流感的流行趋势，构建流行性感冒发病的自回归滑动平均混合模型（autoregressive integrated moving average，ARIMA）并对验证集进行预测，为重庆市流感的防治提供科学依据。**方法：**本研究采用R软件对重庆市2010年1月至2021年6月的流感数据进行ARIMA模型拟合，用2021年7月至2021年12月的数据评价模型拟合效果。**结果：**该流感疾病的流行呈现出明显的季节周期性，周期为1年，发病高峰期为冬季和春季，整体患病率呈上升后有下降趋势，拟合最佳模型为*ARIMA*$(0,1,2)\times(0,1,2)_{12}$，且该模型在预测2021年7月至12月罹患率的均方根误差为10.70，平均绝对百分比误差为70.04%，预测效果较好。**结论：**ARIMA模型在预测重庆市流感发病和流行趋势上有一定的效果，并能对未来的罹患率进行估计，可以为今后的流感疾病防控工作提供参考。

【关键词】流行性感冒；季节乘积自回归移动平均模型；预测

【中图分类号】R511.7　　**【文献标志码】**A　　**【收稿日期】**2023-09-05

Prediction of influenza in Chongqing, China, based on the Autoregressive Integrated Moving Average model

Zou Xiaojiang[1], *Zhao Han*[2], *Wang Qiyin*[1], *Ye Mengliang*[1]

(*1. School of Public Health, Chongqing Medical University; 2. Chongqing Center for Disease Control and Prevention*)

【Abstract】*Objective*: To investigate the trend of influenza prevalence by constructing an Autoregressive Integrated Moving Average (ARIMA) model for influenza and making predictions on the validation set, and to provide a scientific basis for the prevention and control of influenza in Chongqing, China. ***Methods*:** In this study, R software was used for the ARIMA model fitting of influenza data in Chongqing from January 2010 to June 2021, and the data from July to December 2021 were used to evaluate the fitting performance of the model. ***Results*:** The prevalence of this influenza disease presented with a noticeable seasonal pattern, with a yearly cycle and a peak in winter and spring. The overall prevalence rate tended to increase first and then decrease, and the best-fitting model was *ARIMA* $(0,1,2)\times(0,1,2)_{12}$, which had a root mean square error of 10.70 and a mean absolute percentage error of 70.04% in predicting the attack rate in July to December 2021, suggesting that the model had good predictive efficacy. ***Conclusion*:** The ARIMA model has a certain effect in predicting the onset and prevalence trend of influenza in Chongqing and can estimate the attack rate of influenza in the future, which can provide a reference for the prevention and control of influenza in the future.

【Key words】influenza; multiple seasonal autoregressive integrated moving average model; prediction

流行性感冒（简称流感）是一种由流感病毒感染所引起的暴发和流行的呼吸道急性传染病[1]，其流行具有明显的季节性，气象因素、人口规模、流感病毒亚型都会影响其传播[2]。根据世界卫生组织调查估计，全球每年大概有10亿人感染季节性流感病毒，季节性流感相关呼吸系统死亡病例每年29万~64万例[(4.0~8.8)/10万]，其中75岁及以上老年人[(51.3~99.4)/10万]死亡率较高[3]。流感与许多常见疾病存在关联，研究表明，流感病毒感染后3 d内心脏病发作和中风的风险增高[4]，流感与肺炎也存在密切关联，高达50%的肺炎病例与呼吸道病毒有

作者介绍：邹小江，Email：1192690896@qq.com，研究方向：传染病预防与控制。

通信作者：叶孟良，Email：yemengliang@cqmu.edu.cn。

优先出版：https://link.cnki.net/urlid/50.1046.R.20231227.1611.010（2023-12-29）

关[5]。近年来，随着抗生素耐药性问题日趋严重，流感病毒的防控面临着严峻的挑战。2019年3月，世界卫生组织发布了全球性流感防控策略以期进一步加强流感防控。此外，2019年年底一种恶劣的新型冠状病毒开始在世界范围流行，该病毒的入侵给我国带来了巨大的经济和生命财产损失，因其引发的症状与流感类似，具有传播迅速快和致死率高等特点，这也增加了人们对于流感的重视。

随着数学建模的普遍化，越来越多的预测模型用于疾病的预测，其中指数平滑法不适用于长期预测及波动较大的数据预测，对于大流行预测精度会下降；灰色理论模型对于有季节性的传染病的预测效果会受到影响，而自回归滑动平均混合模型（autoregressive integrated moving average，ARIMA）是一种广泛应用于非平稳时间序列分析的线性模型，具有较高的灵活性和参数可解释性，可用于分析时间序列数据中的长期趋势和周期性变动，短期预测效果较好，从国内外相关研究中得知，许多研究者充分证实了用ARIMA模型来预测传染病发病率或发病数是具有一定的可行性与准确性的[6]。ARIMA模型是时间序列预测中常用的模型，在医学领域中也得到了很好的应用[7-9]。因此，本研究采用乘积季节自回归移动平均（seasonal autoregressive integrated moving average，SARIMA）模型对重庆市流感发病率进行预测，为卫生相关部门开展流感疾病防控提供必要的理论依据。

1 资料与方法

1.1 资料来源

流感最新数据来源于重庆市卫生健康委员会官网政务公开疫情信息，包含全市38个区县约3 000万总人口的流感总月度病例数据，人口数据从重庆市统计局获取，与统计年鉴人口总数相对应以获得2010年到2021年重庆市流感每月发病率，本研究数据真实可靠。

1.2 建立ARIMA模型

1.2.1 模型算法 ARIAM模型是一种常用的时间序列模型，由自回归模型（autoregressive model，AR）、差分过程（integrated，I）和移动平均模型（moving average，MA）3部分构成，AR考虑了时间序列的自相关性，即当前值与过去观测值之间存在的线性关系、I用于处理非平稳序列，将数据平稳化可以使得建模效果更佳，MA考虑时间序列的随机波动性即当前值与过去的噪声项之间的关系。ARIMA模型要求数据符合平稳性且非白噪声序列，同时也需要足够长度的历史数据来捕捉时间序列的规律，对于非平稳序列需要进行差分并通过白噪声检验才可用ARIMA模型进行建模。

SARIMA模型以ARIMA模型为基础，又称为ARIMA乘积模型。因为时间序列的内部因素相互之间的关系是十分复杂的，有时候ARIMA模型不能满足建模的需要，此时就需建立SARIMA模型。乘积模型实际上为时间序列的短期相关性和季节效应相乘从而进行拟合，对ARMA和ARMA两者的参数进行相乘，结合d阶的趋势差分使序列平稳化和D阶S步季节差分运算，该模型的呈现形式用$ARIMA(p,d,q)(P,D,Q)s$来表示，若$P=D=Q=0$时，就变成简单的ARIMA模型。

1.2.2 建立模型 使时间序列平稳化：在建模初期，往往发现资料是不平稳的，所以本研究需要经过一系列的方法使它达到平稳性这一目的，获取手段有对数转换或差分等。

ARIMA模型的参数确定：简单的ARIMA模型的3个参数分别为p、d和q，它们的辨别方法主要通过看相关图得出。观察偏自相关函数图确定p值，自相关函数图可确定q值，差分阶数确定d值。在收集整理数据过程中，若发现数据是存在一定的季节趋势，就可以采用模型表达式为$ARIMA(p,d,q)(P,D,Q)s$的乘积季节ARIMA模型，确定P,D,Q三者的参数依旧是通过自相关图。通过上述步骤可以初步确定相应参数的值。

ARIMA模型的建立及预测：根据以上确定的参数来建立ARIMA模型，通过赤池信息量准则（akaike information criterion，AIC）、贝叶斯信息准则（bayesian information criterion，BIC）筛选最佳模型来进行预测。

1.3 统计学方法

本次研究使用Excel 2010整理数据、建立数据库，接着使用R语言4.1.3软件建模。对重庆市2010年1月至2021年12月的发病率监测数据进行数据清理和统计。绘制重庆市流感发病率的时间分布图，以2010年1月至2021年6月的流感监测发病率为训练集，构建乘积季节ARIMA模型并预测2021年7月至2021年12月的发病率，比较预测值与真实值的误差评估模型预测的效果。

2 结 果

2.1 模型建立

2.1.1 流行概况 图1为2010年至2021年重庆市流行性感冒发病率时间分布，主要呈现单峰特点，发病高峰在11月到次年1月，具有明显的季节性特征。从图2的重庆市流感发病率变化趋势可以看出，2018年前重庆市流感发病率都比

较低，2018 年到 2020 年有增加到趋势，在 2019 年末达到最高，之后有下降趋势，但相对 2018 年前发病率依然较高。绘制 2010 年至 2021 年重庆市流感发病率的序列分解图（图 3），可以发现其具有明显的季节性特征且呈现波动上升趋势。

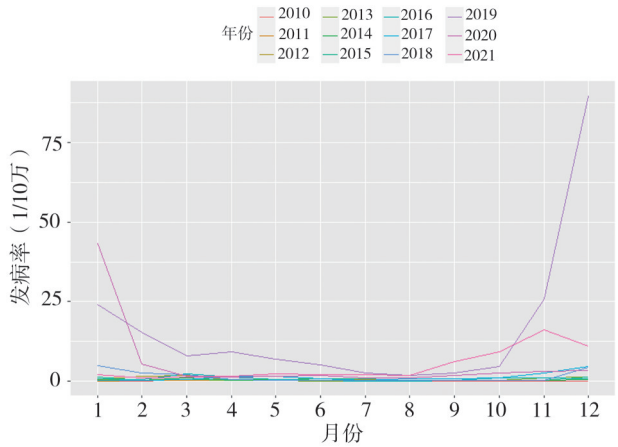


图 1　2010 年至 2021 年重庆市流行性感冒发病率时间分布

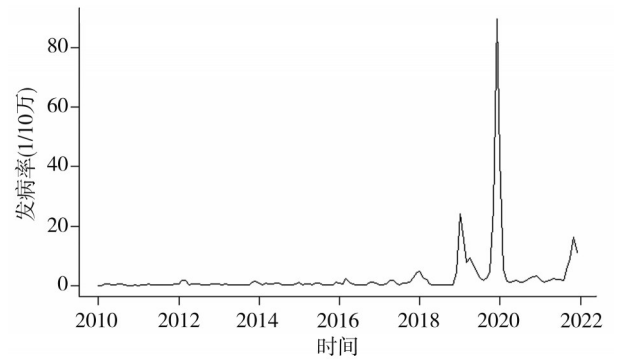


图 2　2010 年至 2021 年重庆市流行性感冒发病率变化趋势

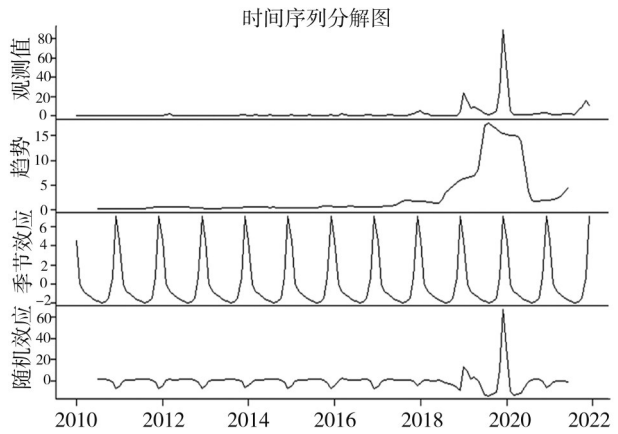


图 3　2010 年至 2021 年重庆市流感发病率序列分解图

2.1.2　季节性差分处理　将 2010 年 1 月到 2021 年 6 月的数据作为训练集，将 2021 年 7 月至 2021 年 12 月的数据作为测试集，考虑到流感的周期性和季节性，对序列进行了一阶逐期差分及长度为 12 阶的季节差分，得到一个基本围绕 0 值波动的平稳序列（图 4），经单位根（augmented dickey-fuller test，ADF）检验，结果显示该序列是平稳的，对差分后的序列进行白噪声检验，延迟 6 阶和 12 阶的 Ljung-Box 检验统计量均小于显著性水平，认为该序列为非白噪声序列，可以用于构建时间序列模型。

2.1.3　构建时间序列模型　考察差分后序列的自相关图和偏自相关图（图 4），对模型的参数进行逐步试阶，对通过了残差白噪声检验和模型参数显著性检验的模型的 AIC 和 BIC 如表 1 所示，进行拟合优度检验，根据 AIC 和 BIC 最小的原则，确定最优模型为 $ARIMA(0,1,2)\times(0,1,2)_{12}$，其 AIC 和 BIC 分别为 872.09 和 886.23，模型参数估计值如表 2 所示，所有估计参数均通过显著性检验。

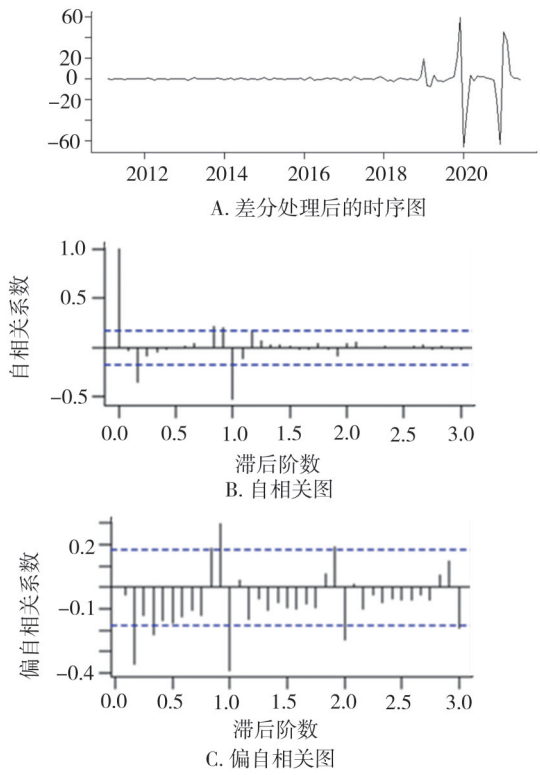


图 4　差分后序列的自相关图和偏自相关图

表 1　模型的拟合优度检验

模型	AIC	BIC
$ARIMA(1,1,2)\times(0,1,1)_{12}$	873.41	887.55
$ARIMA(0,1,4)\times(0,1,1)_{12}$	875.36	892.33
$ARIMA(0,1,4)\times(0,1,2)_{12}$	871.92	891.73
$ARIMA(2,1,1)\times(0,1,2)_{12}$	872.53	889.50
$ARIMA(4,1,1)\times(0,1,2)_{12}$	874.04	896.66
$ARIMA(0,1,2)\times(0,1,2)_{12}$	872.09	886.23
$ARIMA(0,1,2)\times(1,1,1)_{12}$	872.12	886.26

表2 $ARIMA(0,1,2)\times(0,1,2)_{12}$参数检验结果

参数	估计	标准误	t值	P值
MA(1)	−0.258 4	0.086 3	2.994 2	<0.001
MA(2)	−0.602 3	0.089 6	6.719 7	<0.001
SMA(1)	−0.931 8	0.088 7	10.507 2	<0.001
SMA(2)	0.292 7	0.122 8	2.382 5	<0.001

2.2 运用模型进行预测

用$ARIMA(0,1,2)\times(0,1,2)_{12}$模型对2021年7月至12月重庆市流行性感冒发病率进行预测，通过表3和图5可看出，虽然有部分月份的实际值与模型预测的区间是有偏的，但总体而言该模型的预测效果较好，预测值几乎都在95%置信区间内，均方根误差为10.70，平均绝对百分比误差为70.04%，可以用此模型对往后的年份进行预测。

表3 2021年7月至12月重庆市流感发病率的预测

时间	预测值	实际值	差值	95%置信区间
2021年7月	2.23	2.19	−0.04	−11.93~16.38
2021年8月	2.82	1.75	−1.07	−14.80~20.44
2021年9月	3.13	6.26	3.13	−14.60~20.86
2021年10月	4.13	9.14	5.01	−13.71~21.97
2021年11月	11.99	16.16	4.17	−5.96~29.94
2021年12月	36.29	11.11	−25.18	18.23~54.34

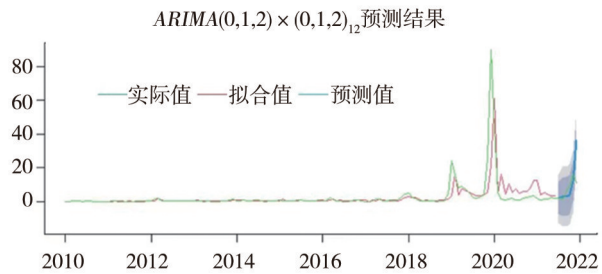


图5 对2021年7月至12月重庆市流感发病率的预测趋势

3 讨 论

随着大量有传染性流感病毒变异株的出现，短时间内流感将仍然存在，近年新型冠状病毒的出现起到了警示作用，我国对于流感的防治工作越来越重视。为了便于掌握疾病的变化趋势、提高卫生服务需求和卫生服务效率、保证居民健康水平，及时为相关疾病卫生防控部门配置医护人员等资源配置提供基础数据，对流感疾病进行科学准确的预测在医学管理范畴和防范决策范畴中是十分必要且关键的，这将是未来卫生部门工作的主要内容之一。

3.1 流行趋势

根据本文研究发现，每年的流感高峰出现在冬春季节，与重庆市以往的流感研究和全国大部分省市的流感研究类似[10-14]。由此可见，本研究具有一定的代表性，可作为防控参考。另外，2019年重庆市流感发病率大幅上升，而后出现一个总体的下降趋势，该趋势的发生可能与同年全国暴发的新型冠状病毒疫情相关，使得该地区的流感监测力度有所增强，医院的就诊率升高。下降趋势可能的原因有：①可能与人群受疫情影响在家里封闭式管理有关；②卫生部门加强了对新冠病毒和流感的预防教育工作。2020年末又出现一个上升的趋势，表明疫情暴发期间流感仍然存在较强的季节性和周期性。

3.2 ARIMA模型拟合与预测的思考

当下数学模型与医学的联系已经相当紧密，许多模型诸如灰色GM预测、指数平滑模型预测、线性回归预测、ARIMA模型[15-17]被用于疾病预测研究并取得诸多成果，其中，在有季节变动和周期性的数据中，ARIMA模型运用得较多，其在短期预测上效果更佳、精确度较高。本次研究通过对2010年1月到2021年6月的流感监测数据进行模型识别、估计、检验和预测，最终显示模型$ARIMA(0,1,2)\times(0,1,2)_{12}$的拟合效果较好，对2021年7月至12月数据进行预测，均方根误差为10.70，预测值与实际值存在差异可能是由于新冠疫情之后，社会防控措施力度加大以及居民对于传染病的防控意识普遍提高，居民更主动采取公共场合戴口罩、勤洗手消毒等措施降低了流感病毒感染的机会，因此实际值与预测值产生较大差异，考虑此因素后该模型的预测效果良好，可用此模型对该地区流感发病率进行预测。然而，观察发现2021年12月拟合度不高，流感发病率实际值与预测值的绝对误差较大，这也说明了此模型会存在一些局限性，即对于大流行之后的预测存在一定的偏差。此外，本研究还存在一些局限性。由于研究仅获得流感病例的月度发病资料，且检测数据可能存在漏报、误报等引起的报告偏倚。而流感的发生和传播是一个复杂的过程，所以需要考虑的因素也比较多，仅使用发病率进行预测会存在一定偏差，因此在未来的建模过程中，可以尝试使用神经网络来拟合、完善模型，同时需要考虑流感的影

响因素，模型预测能力受部环境干扰较大，流感的发生会受到各种疾病防控力度的影响，比如新冠防控措施有效的减少了流感的发生[18]，这些因素在构建预测模型时需要及时调整以提高模型预测的准确度。

综上所述，本研究建立的模型适合运用到本次预测中。从预测的结果可知，未来还会有一定程度的季节性爆发，春冬季前后做好必要的防控措施。该结果对疾控和医院合理配置人力资源具有实际应用价值，卫生防控机构可以根据预测结果采取对应措施，如疾病控制中心可定期通过网络或者宣传栏普及流感防控知识，如提倡人们出门需佩戴好口罩，切断流感的传播途径，保持个人清洁，注意卫生，提高免疫力等。

参 考 文 献

[1] 杨萌萌，龚胜生，黄淑琼，等. 2009—2020年湖北省流行性感冒时空分布特征及预测分析[J]. 公共卫生与预防医学，2023，34(6)：35-39.

Yang MM，Gong SS，Huang SQ，et al.spatial-temporal distribution characteristics and prediction of influenza in Hubei，2009-2020[J]. J Public Health Prev Med，2023，34(6)：35-39.

[2] Ito G，Takazono T，Hosogaya N，et al. Impact of meteorological and demographic factors on the influenza epidemic in Japan：a large observational database study[J]. Scientific Reports，2023，13(1)：13000.

[3] Iuliano AD，Roguski KM，Chang HH，et al. Estimates of global seasonal influenza-associated respiratory mortality：a modelling study[J]. The Lancet，2018，391(10127)：1285-1300.

[4] Smeeth L，Thomas SL，Hall AJ，et al. Risk of myocardial infarction and stroke after acute infection or vaccination[J]. New England Journal of Medicine，2004，351(25)：2611-2618.

[5] Ruuskanen O，Lahti E，Jennings LC，et al. Viral pneumonia[J]. The Lancet，2011，377(9773)：1264-1275.

[6] 王炜翔. 南京市2006～2013年流感流行特征及ARIMA模型的应用[D]. 南京：东南大学，2015.

Wang WX. The analysis of influenza epidemic characteristics and the application of ARIMA model in infections disease prediction in Nanjing from 2006 to 2013[D].Nanjing：Southeast University，2015.

[7] 王 晨，郭 倩，周罗晶. 基于R语言的ARIMA模型对流感样病例发病趋势的预测[J]. 中华疾病控制杂志，2018，22(9)：957-960.

Wang C，Guo Q，Zhou LJ. Forecast of incidence trend of infulenza-like illness by the ARIMA model based on R[J]. Chinese Journal of Disease Control and Prevention，2018，22(9)：957-960.

[8] Meng P，Huang J，Kong D. Prediction of Incidence Trend of Influenza-Like Illness in Wuhan Based on ARIMA Model[J]. Computational and Mathematical Methods in Medicine，2022，2022：6322350.

[9] He Z，Tao H. Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan，China：a nine-year retrospective study[J]. International Journal of Infectious Diseases，2018，74：61-70.

[10] Qi L，Xiong Y，Xiao B，et al. Epidemiological and virological characteristics of influenza in Chongqing，China，2011-2015[J]. PLoS One，2016，11(12)：e0167866.

[11] Meng X，Zhao H，Ou R，et al. Epidemiological and Clinical Characteristics of Influenza Outbreaks Among Children in Chongqing，China[J]. Frontiers in Public Health，2022，10：760746-760746.

[12] 王怡珺，沈福杰，舒 敏，等. 上海市黄浦区流感发病的人群及时空分布特征[J]. 环境与职业医学，2021，38(4)：408-413.

Wang YJ，Shen FJ，Shu M. Demographic and spatial-temporal distributions of influenza cases in Huangpu District，Shanghai[J]. Journal of Environmental and Occupational Medicine，2021，38(4)：408-413.

[13] 傅伟杰，丁 晟，程慧健，等. 2017—2019年江西省流行性感冒时空聚集性分析[J]. 中华疾病控制杂志，2022，26(1)：80-85.

Fu WJ，Ding S，Cheng HJ，et al. Spatiotemporal clustering analysis of influenza in Jiangxi Province from 2017 to 2019[J]. Chinese Journal of Disease Control and Prevention，2022，26(1)：80-85.

[14] Diamond C，Gong H，Sun FY，et al. Regional-based within-year seasonal variations in influenza-related health outcomes across mainland China：a systematic review and spatio-temporal analysis[J]. BMC medicine，2022，20(1)：58.

[15] Guo X，Shen H，Liu S，et al. Predicting the trend of infectious diseases using grey self-memory system model：a case study of the incidence of tuberculosis[J]. Public Health，2021，201：108-114.

[16] Guohun Z，Liping L，Yuebin Z，et al. Forecasting influenza based on autoregressive moving average and holt-winters exponential smoothing models[J]. Journal of Advanced Computational Intelligence and Intelligent Informatics，2021，25(1)：138-144.

[17] Lu Y，Wang S，Wang J，et al. An epidemic avian influenza prediction model based on google trends[J]. Letters in Organic Chemistry，2019，16(4)：303-310.

[18] Wiese AD，Everson J，Grijalva CG. Social distancing measures：evidence of interruption of seasonal influenza activity and early lessons of the SARS-CoV-2 pandemic[J]. Clinical Infectious Diseases，2021，73(1)：e141-e143.

（责任编辑：周一青）