

DOI:10.12138/j.issn.1671-9638.20194086

· 论 著 ·

## 三种时间序列模型预测医院感染发病率的比较

陈越火, 顾翔宇, 于志臻

(复旦大学附属华东医院医院感染管理科, 上海 200040)

**[摘要]** **目的** 比较和评价不同时间序列模型预测医院感染发病率的效果,探索可用于预测医院感染发病率的最佳模型。**方法** 以上海某三级甲等医院 2011—2016 年累计 72 个月的月度医院感染发病率数据作为拟合集构建季节性自回归移动平均模型(ARIMA)、NAR 神经网络模型和 ARIMA-BPNN 组合模型,以 2017 年 1—12 月的月度感染发病率数据作为预测集检验模型的预测效果,评价比较不同模型的预测效果。**结果** 对于拟合集,ARIMA 模型、NAR 神经网络模型和 ARIMA-BPNN 组合模型的 MAPE 分别为 13.00%、14.61%和 11.95%;对预测集,三者的 MAPE 分别为 15.42%、26.31%和 14.87%。**结论** 三种时间序列模型对医院感染发病率均具有较好的预测效果,其中 ARIMA-BPNN 组合模型对拟合和预测该院医院感染发病情况最佳,可为医院决策提供一定的数据支持。

**[关键词]** 医院感染; ARIMA; ARIMA-BPNN 组合模型; NAR 神经网络; 预测

**[中图分类号]** R181.2

## Comparison of three time series models in predicting the incidence of healthcare-associated infection

CHEN Yue-huo, GU Xiang-yu, YU Zhi-zhen (Department of Healthcare-associated Infection Management, Huadong Hospital Affiliated to Fudan University, Shanghai 200040, China)

**[Abstract]** **Objective** To compare and evaluate the effect of different time series models in predicting incidence of healthcare-associated infection(HAI), and explore the best model for predicting incidence of HAI. **Methods** Seasonal autoregressive integrated moving average (ARIMA) model, nonlinear autoregressive neural network (NARNN), and ARIMA-back propagation neural network(ARIMA-BPNN) combination model were constructed based on fitting dataset of monthly HAI incidence from 2011 to 2016 (72 months) in a tertiary first-class hospital in Shanghai, predicting dataset of monthly infection incidence from January to December 2017 were used to test the predictive effect of model, the predictive effect of different models was evaluated and compared. **Results** For the fitting dataset, mean absolute percentage error (MAPE) of ARIMA, NARNN, and ARIMA-BPNN combination model were 13.00%, 14.61%, and 11.95% respectively; and for the predicting dataset, MAPE of ARIMA, NARNN, and ARIMA-BPNN combination model were 15.42%, 26.31%, and 14.87% respectively. **Conclusion** Three time series models can effectively predict the incidence of HAI, of which the ARIMA-BPNN combination model showed the best performance in fitting and predicting the occurrence of HAI in this hospital, and can provide data support for the hospital decision-making.

**[Key words]** healthcare-associated infection; autoregressive integrated moving average; ARIMA; ARIMA-BPNN combination model; nonlinear autoregressive neural network; NARNN; prediction

[收稿日期] 2018-07-17

[作者简介] 陈越火(1985-),男(汉族),浙江省绍兴市人,主管医师,主要从事医院感染管理研究。

[通信作者] 于志臻 E-mail:huadong303@sina.com

医院感染已成为所有医疗机构无法回避的公共卫生问题,随着诊疗技术的不断发展,加之疾病谱的变化和人口老龄化程度的不断提高,给医院感染的预防与控制增加了诸多挑战。系统、持续地对医院感染进行监测和分析,根据监测结果针对性地制定干预措施是保证患者医疗安全,最大限度减少医院感染的重要手段<sup>[1-2]</sup>。而准确、及时、有效地对未来一段时间内医院感染的发生、发展趋势的预测可为医院日常管理、决策提供一定的数据支持。目前医院感染发病率除较多地使用传统的自回归移动平均模型(autoregressive integrated moving average model, ARIMA)预测外<sup>[3-4]</sup>,神经网络模型也开始逐渐应用于医院感染预测领域<sup>[5-6]</sup>。本文使用 ARIMA 模型、NAR 神经网络(nonlinear autoregressive neural network, NARNN)模型和 ARIMA-BPNN(back propagation neural network)串联组合模型分别建立预测模型,比较三者的预测效果,探讨时间序列模型在医院感染领域应用的适用性,探索预测医院感染发病率的最佳模型,为医院的相关决策提供科学依据。

## 1 资料与方法

1.1 数据来源 资料来源于上海某三级甲等医院 2011 年 1 月—2017 年 12 月医院感染综合监测数据,医院感染发病率按照公式:医院感染发病率=同期新发医院感染病例数/观察期间危险人群例数×100%计算,其中新发医院感染病例数据通过临床科室主动报告与医院感染监控专职人员查阅病历相结合的方式获取,医院感染病例依照卫生部 2001 年颁发的《医院感染诊断标准(试行)》诊断,观察期间危险人群例数以同期出院例数替代<sup>[7]</sup>。2011 年 1 月—2016 年 12 月累计 72 个月的月度医院感染发病率数据作为拟合集拟合模型,2017 年 1—12 月的月度感染发病率数据作为预测集检验模型的预测效果。

1.2 模型原理与方法 本次研究通过构建 ARIMA 模型、NARNN 模型及 ARIMA-BPNN 组合模型,比较三种模型的预测效果,以探索可用于预测医院感染发病率的最佳模型。模型构建步骤如下。

1.2.1 ARIMA 模型的建立<sup>[8]</sup> 根据 ARIMA 建模的基本程序,首先对月度发病率数据进行平稳性等序列预处理,根据医院感染发病率存在一定季节性的特征,建立季节性 ARIMA 模型:ARIMA(p, d, q)×(P, D, Q)<sub>12</sub>,其中 p、d、q 分别是简单模型

的自回归阶数、差分阶数和移动平均项,P、D、Q 分别是季节性模型的自回归阶数、差分阶数和移动平均项。根据序列散点图、自相关函数(auto-correlation function, ACF)图、偏自相关函数(partial auto-correlation function, PACF)图等初步确定模型参数,并通过残差检验判断拟合模型是否有效,最后根据赤池信息量(Akaike information criterion, AIC)或贝叶斯信息量(Bayesian information criterion, BIC)选取多个模型的最优模型,其中 AIC 和 BIC 均是模型拟合精度和参数未知个数的加权函数,使 AIC 或 BIC 函数达到最小的模型为相对最优模型。

1.2.2 NARNN 模型的建立<sup>[9]</sup> NAR 神经网络是一种用来模拟时间序列的神经网络,一个典型的 NAR 神经网络主要由输入层、隐含层和输出层及输入和输出的延时构成。网络的输出取决于当前的输入和过去的输出。NARNN 模型方程为: $y(t) = f(y(t-1), \dots, y(t-d))$ ,其中  $y(t)$  是神经网络输出,d 表示延时阶数,f 表示用神经网络实现的非线性函数。根据神经网络构建程序,将拟合集数据随机分成训练样本(80%)、检验样本(10%)和测试样本(10%),采用 Levenberg-Marquard 算法,根据医院感染发病率有一定季节性的特点,设定延时阶数为 12,隐含层神经元个数采用试错方法来确定,分别构建隐含神经元个数从 10 到 100 的模型,由于输入权重和阈值不同,每个模型训练 20 次,依据训练结果的 MSE 和决定系数 R<sup>2</sup> 选取最佳模型。训练阶段使用 open-loop 模式,训练完成后使用 closeloop 函数使改为闭环模式进行预测,预测采用前进递推预测法,即根据拟合集数据预测(t+1)月医院感染发病率,再将(t+1)月感染发病率代入预测模型计算(t+2)月医院感染发病率,以此类推。NAR 神经网络的实现使用 MATLAB R2014b 软件中的神经网络工具箱中的 ntstool。

1.2.3 ARIMA-BPNN 模型的建立 BP 神经网络是一种按误差反向传播训练的多层前馈网络,其基本思想是梯度下降法,利用梯度搜索技术,以期使网络的实际输出值和期望输出值的误差均方值为最小,本研究构建三层 BP 神经网络,输入层 2 个神经元,输出层 1 个神经元,隐含层神经元个数从 10 到 50 分别进行训练。网络训练参数中隐含层选择双曲正切 S 形传递函数(tansig),输出层选择纯线性传递函数(pureline),训练函数选择贝叶斯正规化函数(trainbr)。将前述最优 ARIMA 模型输出的拟合

值和相对应的时间变量(月度)建立一个 2 维矩阵作为 BPNN 的输入变量,将相对应的医院感染发病率实际值数据建立一个 1 维矩阵作为 BPNN 的目标变量,数据用 mapminmax 函数进行归一化处理,通过调整隐含层神经元个数获得组合模型的最优训练结果。

1.2.4 模型验证 使用均方误差(mean-square error, MSE)和相对误差绝对值平均(mean absolute percentage error, MAPE)评价 ARIMA 模型、NARNN 模型和 ARIMA-BPNN 组合模型的拟合与预测效果<sup>[10]</sup>。通过比较 MSE 和 MAPE 的大小

评判模型的优劣,以模型预测效果 MAPE 最小者为最优模型<sup>[11-12]</sup>。

1.3 统计学分析 应用 R 3.4.4 forecast 包构建 ARIMA 模型,使用 MATLAB R2014b 神经网络工具箱构建 NARNN 模型和 ARIMA-BPNN 组合模型。

## 2 结果

2.1 医院感染情况 2011—2017 年本次研究对象的医院感染发病率为 0.99%,分月统计的医院感染发病率为 0.52%~1.77%。见图 1。

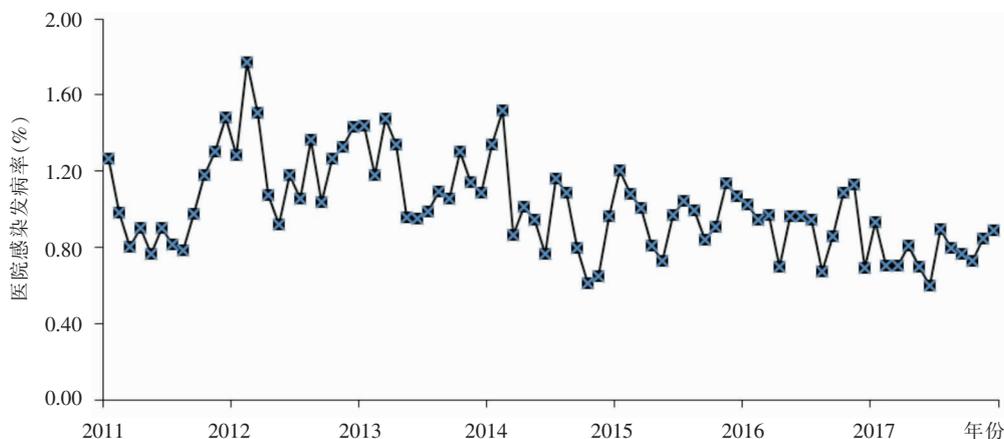


图 1 2011—2017 年月度医院感染发病率变化趋势

Figure 1 Changing trend of monthly incidence of HAI from 2011 to 2017

2.2 ARIMA 模型 首先对拟合集数据进行平稳化处理,经过一阶简单差分和一阶季节性差分后,经增项 DF 单位根(augmented Dickey-Fuller test, ADF)检验显示序列平稳(Dickey-Fuller = -7.091 7,  $P < 0.01$ ),参考 ACF 和 PACF 图的截尾拖尾情况,尝试拟合不同的 ARIMA 模型,根据 AIC 和 BIC 最小和模型简化原则,结合模型拟合效果和残差 QQ 图,确定最优模型为  $ARIMA(0, 1, 1) \times (0, 1, 1)_{12}$ ,模型数学表达式为  $\Delta \Delta_{12} y_t = \epsilon_t - 0.698 2\epsilon_{t-1} - \epsilon_{t-12} + 0.698 2\epsilon_{t-13}$ ,参数检验见表 1, AIC = 1.24, Ljung-Box 检验统计量为 1.1026,  $P = 0.2937$ ,残差序列可认为白噪声。

2.3 NARNN 模型 分析中发现当隐含层神经元个数为 80 时,训练样本的  $MSE = 7.05 \times 10^{-6}$ ,  $R^2 =$

0.999,检验样本的  $MSE = 0.028$ ,  $R^2 = 0.848$ ,测试样本的  $MSE = 0.092$ ,  $R^2 = 0.797$ ,显示模型拟合效果非常理想。误差自相关图显示,误差在 lag 为 0 时最大,其他情况下均未超过可信区间,表示构建的模型可用。模型的总体决定系数  $R^2$  为 0.9064, MSE 为 0.065,拟合值与真实值的误差见图 2。

表 1  $ARIMA(0, 1, 1) \times (0, 1, 1)_{12}$  的参数估计

Table 1 Parameter estimation of  $ARIMA(0, 1, 1) \times (0, 1, 1)_{12}$

变量	估计值	s. e	t	P
MA1	-0.6982	0.0998	-6.99	<0.001
SMA1	-1.0000	0.4413	-2.27	0.013

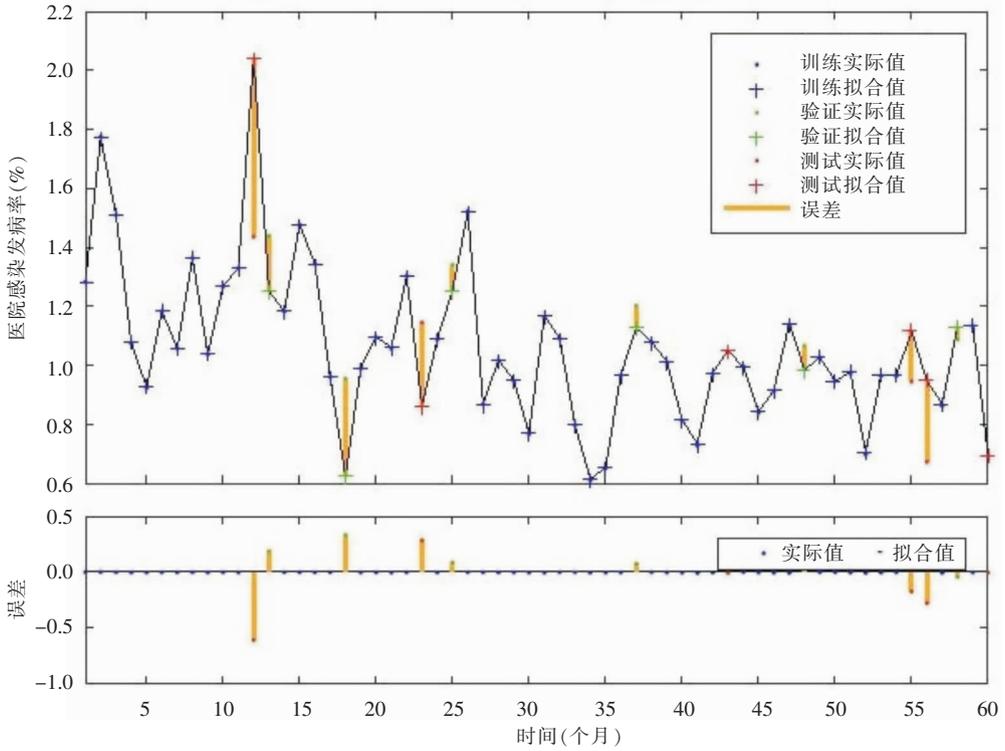


图2 NARNN模型拟合效果误差

Figure 2 Errors of fitting effect of NARNN

2.4 ARIMA-BPNN模型 将前述建立的 ARIMA(0,1,1) × (0,1,1)<sub>12</sub>模型输出的拟合值和相对应的月度时间变量建立一个2维矩阵作为BPNN的输入变量,将相对应的医院感染实际发病率数据

建立一个1维矩阵作为BPNN的目标变量,构造ARIMA-BPNN组合模型。分析中发现当隐含层神经元个数为20时,组合模型MSE为0.025,模型拟合效果较好,发病率拟合值与真实值的比较见图3。

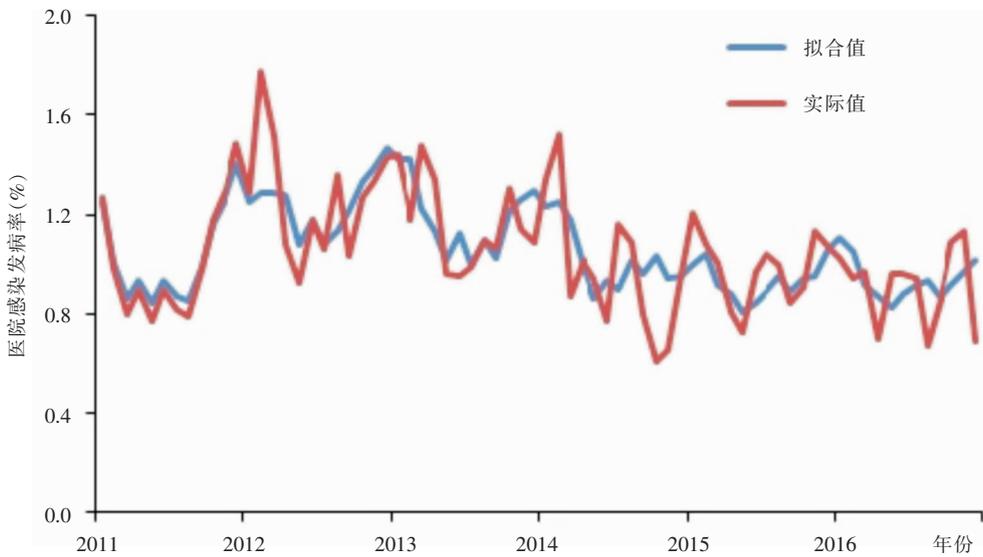


图3 ARIMA-BPNN组合模型拟合效果

Figure 3 Fitting effect of ARIMA-BPNN combination model

2.5 拟合与预测效果 应用 MSE、MPAE 指标比较三种时间序列模型的拟合效果,见表 2,结果显示,ARIMA-BPNN 组合模型的拟合优度最佳,季节性 ARIMA 模型次之,NARNN 模型稍差。以 2017 年 1—12 月的月度感染发病率数据作为预测集,应用 MPAE 指标比较三种模型的预测效果,见表 3,结果显示,季节性 ARIMA 模型、NARNN 模型和 ARIMA-BPNN 组合模型的预测结果与实际值的 MAPE 分别为 15.42%、26.31% 和 14.87%,其中季节性 ARIMA 模型与 ARIMA-BPNN 组合模型

的预测精度接近,后者稍优于前者,两者均明显优于 NARNN 模型。

表 2 三种时间序列模型拟合效果比较

Table 2 Comparison of fitting effect of three time series models

模型	MSE	MPAE(%)
ARIMA	0.030	13.00
NARNN	0.065	14.61
ARIMA-BPNN	0.025	11.95

表 3 2017 年月度医院感染发病率三种模型的预测效果比较

Table 3 Comparison of the performance of three models in predicting the monthly incidence of HAI in 2017

月份	实际值	ARIMA		NARNN		ARIMA-BPNN	
		预测值	MAPE(%)	预测值	MAPE(%)	预测值	MAPE(%)
1	0.93	1.09	16.33	0.89	5.23	0.97	3.96
2	0.71	1.07	52.06	0.93	32.01	0.96	36.36
3	0.71	0.93	32.02	0.85	19.68	0.89	26.40
4	0.81	0.80	0.85	0.75	7.44	0.85	4.65
5	0.70	0.71	1.06	0.88	25.11	0.82	16.86
6	0.60	0.78	30.93	1.06	77.51	0.84	40.65
7	0.89	0.83	7.20	1.05	17.90	0.86	4.02
8	0.80	0.83	4.12	0.64	19.09	0.86	7.93
9	0.77	0.76	1.39	1.07	40.08	0.84	9.27
10	0.73	0.89	21.06	1.20	63.61	0.88	20.15
11	0.84	0.94	11.65	0.89	5.28	0.90	6.76
12	0.89	0.95	6.39	0.87	2.78	0.90	1.42
平均值	-	-	15.42	-	26.31	-	14.87

### 3 讨论

ARIMA 模型是传统的时间序列线性模型,在处理规则数据方面具有诸多优点,但对非线性映射能力较弱,对非线性数据构建模型的效能较低<sup>[13]</sup>;而人工神经网络模型具有良好的非线性映射能力和自适应能力,在解决非线性数据时优势明显<sup>[14]</sup>。研究<sup>[15]</sup>结果表明,ARIMA 模型、人工神经网络模型均已应用于医院感染发病率的预测,取得了良好的预期效果。但医院感染的发生受多种因素共同影响,不同医院因环境不同其医院感染发病率及其发展变化的趋势也不尽相同,对于特定医院的医院感染发病率预测,具体哪种模型预测效果好,需要具体研究探索。

基于医院感染发病率的变化趋势具有线性和非线性的双重特征,本研究在构建 ARIMA、神经网络的基础上,增加 ARIMA 与神经网络组合模型,并对一年的数据进行了短期预测检验,以发现用于医院感染发病率预测的最佳模型。研究显示,基于本研究对象近年来医院感染监测数据,ARIMA-BPNN 组合模型的拟合优度和预测精度均优于 ARIMA 模型和 NARNN 模型,三者拟合优度指标 MAPE 分别为 11.95%、13.00% 和 14.61%,三者的预测精度指标 MAPE 分别为 14.87%、15.42% 和 26.31%,与文献报道的预测模型的精度相似<sup>[11, 16]</sup>,在预测精度要求不是非常高的情况下,三种模型均可应用于预测医院感染发病率的未来变化趋势,其中 ARIMA-BPNN 组合模型的预测效果最佳,ARIMA 模型预测精度稍低于 ARIMA-BPNN 组合模型,而

NARNN 模型由于对未来感染发病率预测的运算数据包含前一期感染发病率,因此对近期的预测效果较好,对远期的预测结果因形成累积误差而较差。

本研究构建的预测模型较好地拟合和预测了医院感染发病率,但由于本研究的医院感染发病率较低<sup>[17-19]</sup>,构建的模型不一定适合于医院感染发病率较高的医院使用。且由于预测模型是依据以往监测数据推测未来情况,随着未来监测数据的不断积累,模型参数也应随之作出调整,以便更好地进行预测分析。另外,医院感染的发生受诸多因素的影响,本研究只使用单一月度医院感染发病率构建模型,其他影响因素未加入到预测模型中,在今后的研究中,考虑尽可能多地收集影响医院感染发病率的各种因素,将之纳入模型,以便进一步提高模型预测精度。

#### [参 考 文 献]

- [1] 胡新华. 多药耐药菌感染情况分析 & 医院感染监测措施的价值研究[J]. 中国医学装备, 2014, 11(10):37-39.
- [2] 周宏, 姜亦虹, 李阳, 等. 176 所医院连续 6 年 ICU 医院感染目标性监测分析[J]. 中国感染控制杂志, 2017, 16(9):810-815.
- [3] 管利华. ARIMA 模型预测医院感染发病状况研究[J]. 实用预防医学, 2013, 20(10): 1247-1249.
- [4] Aldeyab MA, McElnay JC, Scott MG, et al. Hospital antibiotic use and its relationship to age-adjusted comorbidity and alcohol-based hand rub consumption [J]. Epidemiol Infect, 2014, 142(2): 404-408.
- [5] Chen J, Pan QS, Hong WD, et al. Use of an artificial neural network to predict risk factors of nosocomial infection in lung cancer patients[J]. Asian Pac J Cancer Prev, 2014, 15(13): 5349-5353.
- [6] 王高帅, 陈晓娟, 梁进娟. 基于 ARIMA-BPNN 模型对医院感染患病率的预测研究[J]. 中华医院感染学杂志, 2017, 27(2):448-451.
- [7] 中华人民共和国卫生部. 医院感染监测规范: WS/T312-2009 [S]. 北京, 2009.
- [8] 刘峰, 朱妮, 邱琳, 等. ARIMA 乘积季节模型在陕西省手足口病预测中的应用[J]. 中华流行病学杂志, 2016, 37(8): 1117-1120.
- [9] Zhou L, Zhao P, Wu D, et al. Time series model for forecast-

ing the number of new admission inpatients[J]. BMC Med Inform Decis Mak, 2018, 18(1): 39-49.

- [10] 刘思峰. 预测方法与技术[M]. 2 版. 北京: 高等教育出版社, 2015:11-13.
- [11] 陈笑, 王发信, 戚王月, 等. 基于遗传算法的 BP 神经网络模型在地下水埋深预测中的应用——以蒙城县为例[J]. 水利水电技术, 2018, 49(4):1-7.
- [12] Song J, He J. A multistep chaotic model for municipal solid waste generation prediction[J]. Environ Eng Sci, 2014, 31(8): 461-468.
- [13] 严薇荣, 徐勇, 杨小兵, 等. 基于 ARIMA-GRNN 组合模型的传染病发病率预测[J]. 中国卫生统计, 2008, 25(1):82-83.
- [14] 吴伟, 安淑一, 郭军巧, 等. 非线性自回归神经网络在肾综合征出血热流行趋势预测中的应用[J]. 中华流行病学杂志, 2015, 36(12):1394-1396.
- [15] 罗阳, 丁国英, 府伟灵, 等. 季节性时间序列分析在医院感染中的应用[J]. 中华医院感染学杂志, 2005, 15(4):390-392.
- [16] 任宏, 王晔, 陈明亮, 等. 上海市 2005—2012 年猩红热流行特征和发病趋势分析[J]. 中华流行病学杂志, 2013, 34(7):706-710.
- [17] 李娇, 商临萍, 刘玲玉, 等. 山西省某三甲综合医院住院患者医院感染率的调查与分析[J]. 中国感染与化疗杂志, 2016, 16(1):11-14.
- [18] Meneguetti MG, Canini SR, Bellissimo-Rodrigues F, et al. Evaluation of nosocomial infection control programs in health services[J]. Rev Lat Am Enfermagem, 2015, 23(1): 98-105.
- [19] Spatenkova V, Bradac O, Fackova D, et al. Low incidence of multidrug-resistant bacteria and nosocomial infection due to a preventive multimodal nosocomial infection control: a 10-year single centre prospective cohort study in neurocritical care[J]. BMC Neurol, 2018, 18(1): 23-35.

(本文编辑:陈玉华)

**本文引用格式:**陈越火, 顾翔宇, 于志臻. 三种时间序列模型预测医院感染发病率的比较[J]. 中国感染控制杂志, 2019, 18(2): 147-152. DOI:10.12138/j.issn.1671-9638.20194086.

**Cite this article as:** CHEN Yue-huo, GU Xiang-yu, YU Zhi-zhen. Comparison of three time series models in predicting the incidence of healthcare-associated infection[J]. Chin J Infect Control, 2019, 18(2): 147-152. DOI:10.12138/j.issn.1671-9638.20194086.