不同时间序列模型在潍坊市肾综合征出血热预测应用中的比较研究*

郑 良[1] 高 琦[1] 于胜男[1] 石 圆[1] 孙明浩[1] 王志强[2] 李秀君[1△]

【提 要】 **目的** 比较季节自回归移动平均模型(SARIMA)、长短期记忆网络(LSTM)、经验动态建模(EDM)在包含及不包含气象因素的情况下预测潍坊市肾综合征出血热(HFRS)发病的效果,探索最佳预测模型。**方法** 选取 2011 年 1 月至 2017 年 12 月潍坊市 HFRS 月发病率分别构建 SARIMA 模型、单变量 LSTM 模型、单变量 EDM 模型,以及包含气象因素的 SARIMAX 模型、多变量 LSTM 模型、多变量 EDM 模型,对 2018 年 1 月至 2018 年 12 月的月发病率进行预测,并比较各模型的预测效果。**结果** SARIMA 模型的平均绝对误差百分比(MAPE)为 42.17%,SARIMAX 模型未通过参数检验;单变量 LSTM 模型、多变量 LSTM 模型的 MAPE 分别为 48.40%,16.19%;单变量 EDM,多变量 EDM 模型的 MAPE 分别为 55.00%,51.79%。**结论** 包含气象因素的多变量 LSTM 模型对潍坊市 HFRS 发病率预测效果最佳,预测结果可为未来 HFRS 的防控提供参考。

【关键词】 肾综合征出血热 气象因素 SARIMA 模型 LSTM 模型 EDM 模型
【中图分类号】 R195.1 【文献标识码】 A **DOI** 10.11783/j.issn.1002-3674.2024.03.013

Comparison of Different Time Series Models in the Prediction of Hemorrhagic Fever with Renal Syndrome in Weifang

Zheng Liang, Gao Qi, Yu Shengnan, et al (*Department of Epidemiology and Health Statistics, School of Public Health, Shandong University*(250012), *Jinan*)

【**Abstract**】 **Objective** To compare the effects of SARIMA, LSTM and EDM in predicting the incidence of HFRS in Weifang under different circumstances, and explore the best prediction model. **Methods** The monthly incidence of HFRS in Weifang from January 2011 to December 2017 was selected to construct the SARIMA model, univariate LSTM model, univariate EDM model, and SARIMAX model, multivariate LSTM model, and multivariate EDM model including meteorological factors. The monthly incidence from January 2018 to December 2018 was predicted, and the prediction effects of each model were compared. **Results** The MAPE of SARIMA model, univariate LSTM model, multivariate LSTM model, univariate EDM model, multivariate EDM model were 42.17%, 48.40%, 16.19%, 55.00%, 51.79%, respectively. **Conclusion** The multivariate LSTM model including meteorological factors had a good prediction effect on the incidence of HFRS in Weifang, and the prediction results could provide reference for the prevention and control of HFRS.

【**Key words**】 Hemorrhagic fever with renal syndrome; Meteorological factors; SARIMA; LSTM; EDM

肾综合征出血热(hemorrhagic fever with renal syndrome,HFRS)是一种由汉坦病毒引起的自然疫源性疾病,其以啮齿动物为主要传染源,对人类健康造成严重影响。HFRS 广泛流行于欧亚国家,而我国是受其影响严重的国家之一[1]。近年来山东省 HFRS 发病数常居全国前列,其中潍坊市是山东省内较为严重的疫区之一[2]。既往研究表明 HFRS 流行呈明显的季节性分布,气象因素能够在一定程度上通过植被覆盖、鼠类种群密度等影响 HFRS 的发病[3]。然而过去对 HFRS 的预测研究通常只使用发病率,纳入气象因素进行预测的研究相对较少。因此,本研究通过应用长短期记忆网络(long short-term memory,LSTM)[4],季节自回归移动平均模型(seasonal autoregressive integrated moving average model,SARIMA)[5],经验动态建模(empirical dynamic modelling,EDM)[6],并结合气象因素对潍坊市 HFRS 发病趋势进行预测,比较三种模型在包含气象因素和不包含气象因素的情况下的拟合及预测效果,以期为 HFRS 的防控提供数据支撑和科学依据。

资料与方法

1. 资料来源

2011 年至 2018 年潍坊市 HFRS 月发病资料来自中国疾病预防控制信息系统(CISDCP);同时期气象数据来源自中国气象数据网,包括平均温度(℃)、平均相对湿度(%)、累积降雨量(mm)、平均风速(m/s);人口信息来自于《山东省统计年鉴》。

2. 分析方法

(1)季节自回归移动平均模型(SARIMA)

SARIMA 是一种用于时间序列分析的统计模型,在 ARIMA 模型的基础上考虑季节性因素的影响来增强预测的准确性。其表达式 SARIMA$(p,d,q)(P,D,Q)s$ 中,p 为自回归阶数,P 为季节自回归阶数;d 为差

* 基金项目:国家重点研发计划项目(2019YFC1200500;2019YFC1200502)
1. 山东大学公共卫生学院流行病与卫生统计学系(250012)
2. 山东省疾病预防控制中心传染病防治所
△通信作者:李秀君, E-mail:xjli@sdu.edu.cn

分阶数，D 为季节差分阶数；q 为移动平均阶数，Q 为季节移动平均阶数；s 为周期。SARIMAX 则是将多元线性回归与 SARIMA 模型相结合，在 SARIMA 模型的基础上同时考虑外生变量的影响[5]。建模过程通常包括确定合适的阶数，使用样本数据进行参数估计，构建最优 SARIMA 模型，然后加入外生变量分别构建 SARIMAX模型并根据赤池信息量准则（Akaike information criterion，AIC）确定最优模型。

（2）长短期记忆网络（LSTM）

LSTM 是循环神经网络（rerrent neural network，RNN）的变型，它在解决传统 RNN 面临的长期依赖问题上具有显著的优势[7]。其核心组件是 LSTM 单元，每个单元由记忆细胞和三个门组成，包括遗忘门、输入门和输出门。因此 LSTM 可以选择性地遗忘和保留信息，从而更好地解决 RNN 因长期依赖历史信息而导致的梯度消失和梯度爆炸的问题。多变量 LSTM 模型[4]是传统单变量 LSTM 模型的扩展，在 LSTM 的基础上将输入序列扩展为具有多个特征或维度的向量。在包含多个变量的复杂系统中，多变量 LSTM 模型能够同时考虑多个变量之间的相互影响从而更好地捕捉时间序列的特征。

（3）经验动态建模（EDM）

EDM 是一种基于相空间重构理论的用于研究非线性系统的方法，它能够通过时间序列滞后坐标嵌入法直接从原始时间序列数据恢复原系统动力模式[6,8]。其最大的特点是抛弃传统数据分析中的公式化方法，仅仅从时间序列中重构动态系统的行为。根据 Takens 嵌入定理，对于时间无限长且无噪声的 d 维吸引子的时间序列 $\{x(m)\}$，都可以通过嵌入重构一个 n 维（$n\geq 2d+1$）的拓扑不变的相空间[9]，即重构一个与原动力系统等价的相空间，只需考察一个维度的分量，并将它在某些时间延迟点上的测量作为新的维度，而通过这种滞后时间坐标嵌入重建所得到的影子流形则保留了原动力系统的基本特征[10]。简言之，当时间序列足够长时可以在高维空间重建原系统动力模式。

如图 1，假设真实动力系统是由 X、Y、Z 三个相互影响的变量组成的非线性动力系统。单变量 EDM 模型通过单一嵌入的方法，将时间序列 $X(t)$ 及其滞后作为新的维度重构相空间以完成对 X 未来趋势的预测；多变量 EDM 模型则是通过多元嵌入，将 $X(t)$、$Y(t)$ 及其滞后一起纳入作为新的维度重构相空间以完成对 X 未来趋势的预测[11]。

3. 预测模型构建

选取 2011 年 1 月至 2017 年 12 月潍坊市 HFRS 月发病率分别建立 SARIMA 模型、单变量 LSTM 模型、单变量 EDM 模型，以及包含气象因素的 SARIMAX 模型、多变量 LSTM 模型、多变量 EDM 模型，为使模型之间能够进行比较并且提高预测精度，以滚动预测方法对 2018 年 1 月至 2018 年 12 月的月发病率进行预测。最后采用平均绝对误差百分比（mean absolute percentage error，MAPE）、平均绝对误差（mean absolute error，MAE）、均方根误差（root mean squared error，RMSE）作为预测模型效果的评价指标。

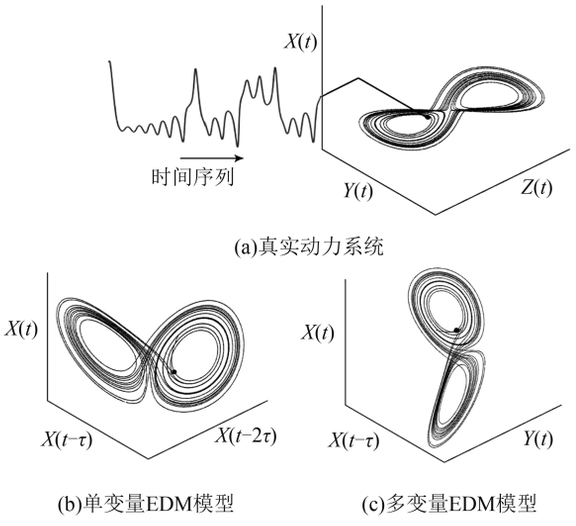

图 1　真实动力系统、单变量 EDM、多变量 EDM 模型对比示意图

4. 分析软件

模型的建立均使用 R 4.0.5 软件。检验水准 $\alpha=0.05$。

结　果

1. 肾综合征出血热发病概述

2011-2018 年潍坊市累积报告 HFRS 发病 2236 例，月平均发病率为 0.25/10 万。对发病率原始时间序列进行分解，如图 2，潍坊市 HFRS 疫情整体形势稳定，在 2012 年出现小高峰；季节性显著，呈双峰分布，存在较为平缓的春峰以及陡峭的冬峰。

2. 相关性分析

2011-2018 年潍坊市气象因素与发病率间的分布均非正态，因此使用 Spearman 相关分析，结果见表 1。月平均温度与月总降水与 HFRS 月发病率之间的相关系数分别为 -0.40，-0.32，且均具有统计学意义（$P<0.05$），因此将温度以及降水纳入后面的预测模型探究。

3. SARIMA 模型与 SARIMAX 模型

对 2011-2017 年的月发病率时间序列进行单位根检验，为非平稳序列（$P=0.289$），对其进行 1 阶差分以及季节差分，得到平稳时间序列（$P<0.05$）。绘制自相关及偏自相关图，确定 p，q 的大概范围，并对参数进行逐一尝试，最终确定 ARIMA$(0,1,2)\times(2,1,0)_{12}$ 为最优 SARIMA 模型（AIC = -71.46）。在 ARIMA

$(0,1,2)\times(2,1,0)_{12}$模型的基础上，分别将温度、降水作为外生变量建立 SARIMAX 模型。结果如表 2，下列 SARIMAX 模型中气象变量作为模型参数均无统计学意义，因此舍弃 SARIMAX 模型。

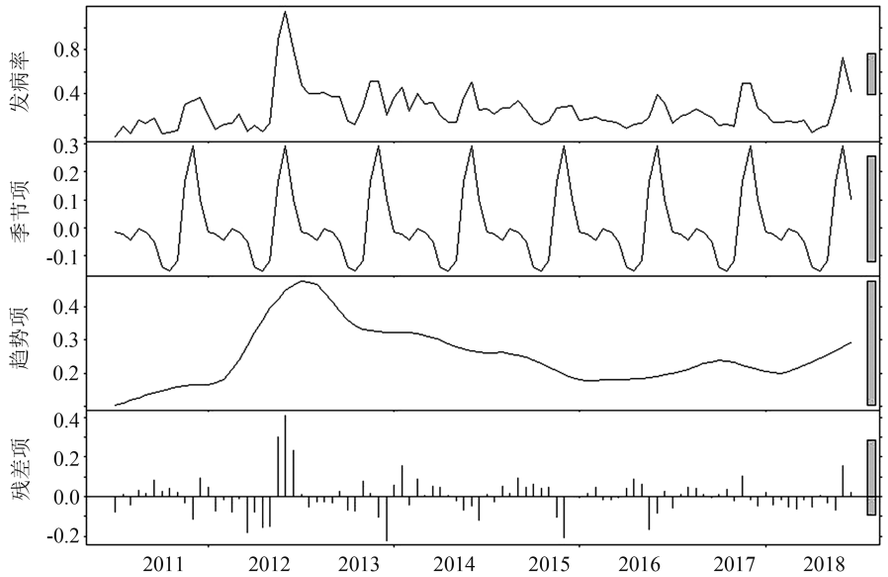

图 2 潍坊市 2011-2018 年 HFRS 月发病率时间序列分解图

表 1 气象因素与 HFRS 发病率的 Spearman 相关性分析

	温度	降水	相对湿度	风速
降水	0.67*			
相对湿度	0.45*	0.61*		
风速	0.09	0.03	−0.50*	
HFRS 发病率	−0.40*	−0.32*	−0.17	0.03

*：$P<0.05$。

4. 单变量 LSTM 模型与多变量 LSTM 模型

对于单变量 LSTM，利用 HFRS 月发病率作为输入序列。而对于多变量 LSTM，分别以三种变量组合作为输入序列：一是发病率与温度；二是发病率与降水；三是发病率、温度与降水。考虑 HFRS 时间序列的季节性和趋势性，将滑动窗口设置为 12，并对模型参数进行网格搜索，以找到适合每个输入配置的最佳参数。每组模型参数的训练都进行 10 次 100 个迭代的尝试，最后选择不同输入配置下的最优模型，结果如表 3。

表 2 SARIMA 模型与包含不同气象因素的 SARIMAX 模型的参数检验及预测结果

模型	参数	回归系数检验		Ljung-Box 检验		AIC	MAPE(%)	
		系数	P	χ^2	P		拟合	预测
$ARIMA(0,1,2)\times(2,1,0)_{12}$				0.001	0.98	−71.46	40.39	42.17
$ARIMA(0,1,2)\times(2,1,0)_{12}$+温度	温度	−0.0047	0.31	0.002	0.96	−69.71	40.30	43.15
$ARIMA(0,1,2)\times(2,1,0)_{12}$+降水	降水	−0.0001	0.38	0.002	0.97	−69.66	40.47	42.74
$ARIMA(0,1,2)\times(2,1,0)_{12}$+温度+降水	温度	−0.0050	0.35	0.003	0.96	−67.96	40.54	43.96
	降水	−0.0001	0.40					

表 3 单变量 LSTM 模型与包含不同气象变量的多变量 LSTM 模型拟合及预测结果

模型	气象变量	拟合			预测		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE
单变量 LSTM		0.06	0.05	23.19	0.11	0.08	48.40
多变量 LSTM	月平均温度	0.02	0.01	6.97	0.03	0.02	16.19
	月降水量	0.01	0.01	5.39	0.04	0.03	16.82
	月平均温度+月降水量	0.01	0.01	4.81	0.04	0.03	17.91

5. 单变量 EDM 模型与多变量 EDM 模型

对于单变量 EDM，以 2011-2017 年月发病率时间序列及其滞后进行相空间重构，对 2018 年月发病率进行预测。对于多变量 EDM，分别以三种变量组合进行相空间重构：一是发病率与温度；二是发病率与降水；三是发病率、温度与降水。为了防止构建扭曲的相空间，对发病率时间序列进行 1 阶差分使其平稳。每种变量组合又有不同的嵌入方法，对所有可能的模型进行尝试，最后每种变量组合选取一个拟合以及预测效果最佳的模型，结果如表 4。

表 4　单变量 EDM 模型与包含不同气象变量的多变量 EDM 模型拟合及预测结果

模型	气象变量	拟合			预测		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE
单变量 EDM		0.15	0.10	49.13	0.13	0.10	55.00
多变量 EDM	月平均温度	0.17	0.11	47.86	0.09	0.08	51.79
	月降水量	0.17	0.12	55.16	0.13	0.10	62.40
	月平均温度+月降水量	0.15	0.09	42.85	0.10	0.08	56.85

6. 预测模型比较

结合上文，分别选择各组表现最佳的预测模型进行比较。考虑到 SARIMAX 模型中的气象变量均未通过参数检验，我们忽略 SARIMAX 模型；最优多变量 LSTM 模型为纳入温度与发病率作为输入的 LSTM 模型；最优多变量 EDM 模型为纳入温度与发病率作为输入的 EDM 模型。各模型预测结果见图 3，预测性能比较见表 5。

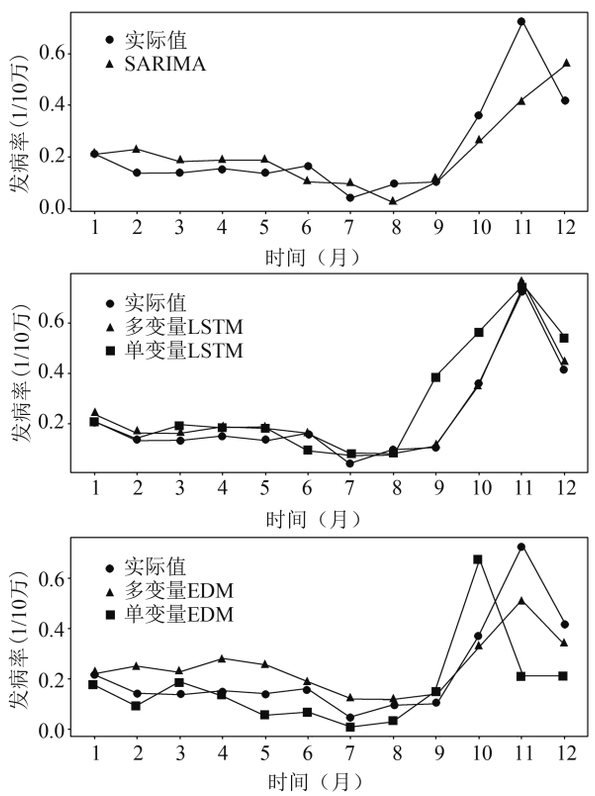

图 3　各模型 HFRS 月发病率预测结果

讨　论

潍坊市作为国内 HFRS 高发区之一，近年来尽管取得了一定的疫情控制成果，但仍面临着重要的公共卫生挑战。在过去的研究中已发现 HFRS 的发病与气象因素密切相关，因此研究包含气象因素的预测模型是否能提高预测精度对完善疫情预警系统具有现实意义。

研究结果显示建立的 SARIMAX 模型中纳入的气象变量均无统计学意义，这或许是因为 SARIMA 在进行差分后已经解决了趋势性与季节性的问题，再将气象因素纳入模型容易导致过拟合。且气象因素与 HFRS 之间存在复杂的非线性关系[12]，而 SARIMAX 是时间序列模型与多元线性回归的结合，因此无法捕捉气象因素对 HFRS 的非线性影响。通过对不同模型进行比较，结果发现在短期预测潍坊市 HFRS 发病率时，除 SARIMAX 模型以外包含气象变量的模型的预测效果都较之前有所提升，其中结合气象变量的多变量 LSTM 模型预测效果最佳，其次为 SARIMA 模型，再次是单变量 LSTM 模型。这说明纳入气象因素有利于提高模型的预测精度，其中 LSTM 作为一种高级的循环神经网络能够进行深度学习并捕获变量之间的非线性依赖，从而能够较好地捕捉气象因素与 HFRS 间的非线性关系[7,13]。而利用 EDM 进行时间序列预测需要使用足够长的时间序列才能重建原系统动力模式[14]，本研究中纳入时间序列较短，因此 EDM 模型在对 HFRS 的预测中表现相对欠佳。

本研究局限性在于发病数据来自传染病报告信息管理系统，可能存在误报情况；以及考虑到在研究大范围区域时，气象因素的影响可能会被稀释，本研究仅针对潍坊市进行了分析，因此研究结论可能不具有普遍性。

综上所述，本研究运用结合气象变量的多变量 LSTM 模型较为准确地预测了潍坊市 HFRS 的发病趋势，希望能为当地制定更具针对性的疫情防控方案提

表 5　预测模型拟合以及预测效果比较

模型		拟合			预测		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE
单变量模型	SARIMA	0.12	0.08	40.39	0.11	0.08	42.17
	单变量 LSTM	0.06	0.05	23.19	0.11	0.08	48.40
	单变量 EDM	0.15	0.10	49.13	0.13	0.10	55.00
多变量模型	多变量 LSTM	0.02	0.01	6.97	0.03	0.02	16.19
	多变量 EDM	0.17	0.11	47.86	0.09	0.08	51.79

供参考。后续将拓展研究地域并在不同时间尺度上对 HFRS 发病趋势进行深入探究，以期为 HFRS 预警系统的完善提供科学依据。

参 考 文 献

[1] He J, Wang Y, Wei X, et al. Spatial-temporal dynamics and time series prediction of HFRS in mainland China: A long-term retrospective study. J Med Virol, 2023, 95(1): e28269.

[2] 范俊杰, 于绍起, 王怡, 等. 潍坊市肾综合征出血热空间分布特征研究. 现代预防医学, 2020, 47(3): 412-418.

[3] Cao L, Huo X, Xiang J, et al. Interactions and marginal effects of meteorological factors on haemorrhagic fever with renal syndrome in different climate zones: Evidence from 254 cities of China. Sci Total Environ, 2020, 721: 137564.

[4] Yang E, Zhang H, Guo X, et al. A multivariate multi-step LSTM forecasting model for tuberculosis incidence with model explanation in Liaoning Province, China. BMC Infect Dis, 2022, 22: 490.

[5] 景钦隆, 吴琦琳, 鲁影, 等. 手足口病流行时间序列模型及其与气象因素联合预测研究. 中国卫生统计, 2020, 37(3): 354-358.

[6] Ye H, Sugihara G. Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality. Science, 2016, 353(6302): 922-925.

[7] Zhu H, Chen S, Liang R, et al. Study of the influence of meteorological factors on HFMD and prediction based on the LSTM algorithm in Fuzhou, China. BMC Infectious Diseases, 2023, 23(1): 299.

[8] Sugihara G, May R, Ye H, et al. Detecting causality in complex ecosystems. Science, 2012, 338(6106): 496-500.

[9] Takens F. Detecting strange attractors in turbulence//RAND D, Dynamical Systems in turbulence, 1981: 366-381.

[10] 王丹雨, 朱媛君, 杨晓晖. 收敛交叉映射方法及其在生态学中的应用. 应用生态学报, 2021, 32(12): 4539-4548.

[11] Deyle ER, Bouffard D, Frossard V, et al. A hybrid empirical and parametric approach for managing ecosystem complexity: Water quality in Lake Geneva under nonstationary futures. Proc Natl Acad Sci USA, 2022, 119(26): e2102466119.

[12] Lv CL, Tian Y, Qiu Y, et al. Dual seasonal pattern for hemorrhagic fever with renal syndrome and its potential determinants in China. Science of The Total Environment, 2023, 859: 160339.

[13] Zhang R, Song H, Chen Q, et al. Comparison of ARIMA and LSTM for prediction of hemorrhagic fever at different time scales in China. PLoS One, 2022, 17(1): e0262009.

[14] Johnson B, Munch SB. An empirical dynamic modeling framework for missing or irregular samples. Ecological Modelling, 2022, 468: 109948.

（责任编辑：郭海强）

（上接第 392 页）

[4] 万崇华,杨铮,李晓梅.慢性病患者生命质量测评手册.北京:科学出版社,2019.

[5] 万崇华,巫小玉,刘钰曦,等.慢性病患者生命质量测定量表体系第2版QLICD(V2.0)研究与应用现状.广东医科大学学报,2022,40(3):243-249.

[6] 范引光,张文慧,陈国平.统计分析中检验方法的选择.中华疾病控制杂志,2010,14(1):86.

[7] 颜艳,王彤.医学统计学·第5版.北京:人民卫生出版社,2020.

[8] 赵进文.经济计量诊断学.天津:天津人民出版社,2000.

[9] 朱钰,郑屹然,尹默.统计学意义下的多重共线性检验方法.统计与决策,2020,36(7):34-36.

[10] 谢秋山,李百超.村委会产生模式感知对村民政治参与权利认知影响研究:基于CGSS2013数据的实证分析.福建行政学院学报,2018(5):54-64.

[11] 王兴华,王大华,申继亮.社会支持对老年人抑郁情绪的影响研究.中国临床心理学杂志,2006(1):73-74+90.

[12] 刘超,孙晓晶,张晓妍,等.2型糖尿病患者心理弹性与生存质量的关系研究.医学与哲学(B),2012,33(12):52-53+66.

[13] 叶艳,范方,陈世键,等.心理弹性、负性生活事件和抑郁症状的关系:钢化效应和敏化效应.心理科学,2014,37(6):1502-1508.

[14] 袁飞.慢性病老年人心理弹性初步研究.重庆:重庆师范大学,2019.

[15] 阙霜,曾雁冰,方亚.基于logistic回归与决策树模型的社会资本对老年人自评健康的影响研究.中国卫生统计,2022,39(2):186-191.

[16] Cao J, Rammohan A. Social capital and healthy ageing in Indonesia. BMC Public Health, 2016, 16(1): 631.

[17] 窦蕾,周萍,李晨,等.上海市康复医院康复资源与服务开展情况调查研究.中国康复医学杂志,2017,32(1):90-93.

[18] 刘倩汝,王梦娜,耿力.我国医养结合养老背景下老年康复护理模式研究进展.护理学杂志,2022,37(5):20-23.

（责任编辑：郭海强）