技术方法 DOI:10.13406/j.cnki.cyxb.001308

痛风月发病趋势预测及预测模型效果探讨

胡 敏,杜成凤,唐晓君,邓 丹
（重庆医科大学公共卫生与管理学院卫生统计与信息管理教研室,重庆 400016）

【摘 要】**目的**:选择最优时间序列模型进行重庆市痛风月发病趋势预测。**方法**:回顾性收集重庆市某三甲医院 2011 年 1 月至 2014 年 12 月痛风月发病数据,并采用指数平滑法、ARIMA(p,d,q)模型、GM(1,1)模型和 Markov 模型对其发病趋势进行拟合,根据平均百分绝对误差的大小判断模型优劣,选择最优时间序列模型进行痛风月发病趋势预测。**结果**:2012 年 9 月后痛风月发病人数激增,并呈现一定的周期性。但就整个观察期而言,序列波动明显,周期性不明显。4 种时间序列模型分析结果显示:实际发病数在各模型预测值 95%的可信范围内波动,但不同模型预测结果均存在一定程度偏离。通过计算可知:ARIMA(1,1,0)、GM(1,1)和指数平滑法的平均百分绝对误差分别为 24.72、109.21 和 24.95。因此 GM(1,1)模型预测效果偏离最大,指数平滑法次之,ARIMA(1,1,0)模型拟合效果最贴近实际。**结论**:GM(1,1)模型拟合效果最差,ARIMA(1,1,0)模型拟合效果最好,建议用 ARIMA(1,1,0)模型进行未来痛风发病预测。

【关键字】痛风;时间序列分析;趋势预测;模型选择
【中图分类号】R589.7 【文献标志码】A 【收稿日期】2016-11-16

Trends prediction of monthly gout incidence and comparison of model performance among four time series models

Hu Min, Du Chengfeng, Tang Xiaojun, Deng Dan
(*Teaching and Research Section of Health Statistics and Information Management, School of Public Health and Management, Chongqing Medical University*)

【Abstract】***Objective***: To choose the optimal time series model for monthly gout incidence prediction. ***Methods***: Monthly gout incidence data were collected retrospectively from an affiliated hospital of Chongqing Medical University from January 2011 to December 2014. And exponential smoothing method, ARIMA(p,d,q), GM(1,1) and Markov model were adopted to fit the trend of monthly gout incidence, respectively. The efficiency comparison between different models was judged by mean absolute percentage error, and model with least mean absolute percentage error was used for future prediction. ***Results***: Monthly gout incidence increased greatly, and cyclicality was observed after September 2012. But only sequence fluctuation was observed on the whole observation period. Results of trend prediction among four time series models were in consistent with each other, but their efficacy were different. The maximum and minimum mean absolute percentage error was in GM(1,1) model and ARIMA(1,1,0) model, respectively. Thus fitting results of ARIMA (1,1,0) model was mostly close to the actual situation. ***Conclusion***: The ARIMA(1,1,0) model have the best imitative effect. Hence, ARIMA(1,1,0) model should be used for future monthly gout incidence prediction.

【Key words】gout; time series analysis; trends prediction; model selection

痛风[1-5]是单钠尿酸盐沉积所致疾病,与嘌呤代谢紊乱所致高尿酸血症直接相关,其典型临床表现包括急性关节炎、痛风石、关节畸形和关节功能障碍等。痛风既往属于“罕见病”“帝王病”,但随全人类饮食结构改善和生活水平提高,全球痛风患病明显上升,使痛风逐渐成为一种常见病、多发病。目前全球共有 1.2 亿痛风患者,较 15 年前增加 15~30 倍[6]。据 2004 年中国疾病与健康调查报告指出[6-7]:目前我国已有超过 7 500 万痛风患者,并且正以 0.97%的增长率继续增加,使痛风成为影响我国居民健康的

作者介绍:胡 敏,Email:359703638@qq.com,研究方向:人群健康与统计决策。
通信作者:邓 丹,Email:ddlinger@126.com。
基金项目:国际合作资助项目(美国杜克大学全国健康研究中心,编号:x7383)。
优先出版:http://kns.cnki.net/kcms/detail/50.1046.r.20170611.1337.006.html (2017-06-09)

第二大慢性代谢性疾病。既往对于痛风的研究都是基于其发病机制[8]、治疗方案[9]、并发症[10-12]等，尚未见对其发病趋势的预测研究。本研究收集了重庆市某三甲医院 2011 年 1 月至 2015 年 5 月的痛风月发病数据，通过 4 种时间序列分析方法建立预测模型，并比较不同模型预测效果，选出最优模型进行未来发病趋势预测，以期为重庆市痛风防治提供依据。

1 资料和方法

1.1 资料来源

回顾性收集了重庆市某三甲医院 2011 年 1 月至 2015 年 5 月出院诊断为痛风及其并发症的 2 552 例患者资料。其中，男性 1 937 例(75.90%)，年龄分布 15~97 岁，已婚者 2 447 例(95.89%)，医疗保险患者 2 308 例(90.44%)，2 464 例(96.44%)痛风患者存在合并症，其中 56.58%合并高血压，27.04%合并糖尿病，23.82%合并脂代谢异常，31.62%合并心脑血管疾病。

1.2 时间序列分析方法

1.2.1 指数平滑法 指数平滑法通过对远期和近期数据差异性赋权来实现历史数据的加权修匀，从而提取序列的确定性信息。预测模型通式为：

$$S_t=\partial x_t+(1-\partial)S_{t-1}$$

其中，∂ 为平滑系数，$\partial\in(0,1)$，x_t 为第 t 期实际观察值，S_t、S_{t-1} 分别为第 t、$t-1$ 期平滑值($t>0$)。初始平滑值 S_0 的确定方法有 3 种，本研究中令序列初值 x_1 为初始平滑值。平滑系数 ∂ 的取值与数据趋势和研究目的有关。本研究数据序列长期变化趋势不明显且研究目的主要是进行未来预测，因此 ∂ 取值为 0.1~0.4。

1.2.2 Markov 模型 Markov 模型是具有无后效性的随机数列模型，具体可表述为：时刻 t 处于状态 i 的随机过程$\{X_t\}$，$t\in T$，在下一时刻处于状态 j 的概率 $P_{ij}(t)$是一固定值。也即：

$$P\{X_{t+1}=j\,|\,X_0=k_0,X_1=k_1,\ldots,X_{t-1}=k_{t-1},X_t=i\}=P\{X_{t+1}=j\,|\,X_t=i\}$$

Markov 模型的状态划分和转移概率矩阵的确定应有专业上的依据。本研究根据月痛风发病人数的构成将其划分为 5 个状态，通过统计各状态的频数得到初始状态概率向量 a_0，并将其一步转移概率矩阵记 $P(t)$为：

$$P(t)=\begin{vmatrix}P_{00}(t) & P_{01}(t)\ldots P_{0n}(t)\\ P_{10}(t) & P_{11}(t)\ldots P_{1n}(t)\\ \ldots\ldots & \\ P_{n0}(t) & P_{n1}(t)\ldots P_{nn}(t)\end{vmatrix}=\begin{vmatrix}0.286 & 0.714 & 0 & 0 & 0\\ 0.333 & 0.5 & 0.083 & 0.083 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0.667 & 0.333\\ 0 & 0 & 0 & 0.214 & 0.786\end{vmatrix}$$

根据公式 $a_n=a_0p_n$ 对未来第 n 个时点的发病率进行预测，选取 a_n 中概率最大时点所属状态作为最后预测值。

1.2.3 ARIMA(p,d,q)模型 ARIMA(p,d,q)模型的建模是在 SAS/ETS 板块中进行，该方法建模是针对平稳非白噪声序列进行。建模流程包括了模型识别与定阶、参数估计和显著性检验。值得注意的是：如果模型无效或参数检验不显著时，应当重复上述步骤直到模型显著为止。当存在多个可选择的模型时，应选择所有参数都显著且具有最小 AIC 和 BIC 值的模型作为终模型。

1.2.4 GM(1,1)预测模型 GM(1,1)模型建模在 SAS/IML 板块中进行，建模流程如下：

首先，生成原始数列 $x^{(0)}$的一次累加数列 $x^{(1)}=(x^{(1)}_{(1)},x^{(1)}_{(2)},\ldots x^{(1)}_{(n)})$，并利用加权法获得一次累加数列 $x^{(1)}$的邻值生成数列 $z^{(1)}=(x^{(1)}_{(2)},x^{(1)}_{(3)},\ldots x^{(1)}_{(n)})$；

其次，根据一次累加生成数列 $x^{(1)}$的灰导数 $d(x)=x^{(0)}(k)=x^{(1)}(k)-x^{(1)}(k-1)$，得到 $z^{(1)}(k)=\partial x^{(1)}(k)+(1-\partial)x^{(1)}(k-1)$，于是定义 GM(1,1)的灰微分方程模型为：$d(k)+az^{(1)}(k)=b$

即 $x^{(0)}(k)+az^{(1)}(k)=b$

其中，$x^{(0)}(k)$为灰导数，a 为发展系数，$z^{(1)}(k)$为白化背景值。由于$\frac{\Delta x^{(0)}}{\Delta t}$涉及累加数列 $x^{(1)}$前后两个时刻的值，因此用两时刻的平均值代替 $z^{(1)}_{(i)}$得到模型表达式为：

$$Y=\begin{cases}x^{(0)}_{(2)}\\ x^{(0)}_{(3)}\\ \cdots\cdots\\ x^{(0)}_{(N)}\end{cases}\qquad \mu=\begin{bmatrix}a\\ b\end{bmatrix}\qquad B=\begin{vmatrix}-\frac{1}{2}[x(1)_{(1)}+x(1)_{(2)}]\\ -\frac{1}{2}[x(1)_{(2)}+x(1)_{(3)}]\\ \cdots\cdots\\ -\frac{1}{2}[x(1)_{(n-1)}+x(1)_{(n)}]\end{vmatrix}$$

最后，根据最小二乘法得到参数 a,b 的估计值：

$$\hat{\mu}=\begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix}=(B^TB)^{-1}B^TY$$

并得到 GM(1,1)一阶线性微分方程的解为：

$$\hat{x}^{(1)}(k+1)=(x^{(0)}_{(1)}-\frac{b}{a})e^{-a(k)}+\frac{b}{a},k=1,2,\ldots,n$$

1.2.5 不同模型预测效果的比较及回代预测 采用上述 4 种时间序列模型对 2011 年 1 月至 2014 年 12 月痛风月发病数据建模，并以 2015 年 1 至 5 月的实际发病人数验证模型的外推效果，并通过平均百分绝对误差(mean absolute percentage error，MAPE)的大小选出最优模型。此外，本研究还将 2015 年 1 至 5 月的数据代入建立的最优模型中进行回代预测。平均百分绝对误差的计算公式为：

$$\text{MAPE}=\frac{\sum\left|\frac{Y_i-\overline{Y}}{Y}\times100\right|}{n}$$

其中，Y_i 为实测值，$\overline{Y}$ 为预测值，n 为样本例数。

2 结 果

2.1 痛风发病序列描述

图 1 为 2011 年 1 月至 2014 年 12 月痛风月发病时序图。由该图可知：2012 年 9 月以前痛风发病人数变化不明显，但该时间段后，痛风发病人数激增，并呈现一定周期性，以 3 月人数为最低。但就整个观察期而言，序列波动较为明显，但周期性不明显。

2.2 时间序列模型分析结果

2.2.1 指数平滑法结果 痛风月发病序列存在一定程度波动，但长期趋势变化不明显。根据经验判断法，平滑系数的取值区间为 0.1~0.4。表 1 为不同平滑系数取值下的各类误差结果。由表 1 可知，平滑系数取值 0.30 时，模型预测误差最小。平滑系数取值小，导致序列的历史信息提取较多，模型反应速度慢，不能及时追踪新数据的变化趋势，因此笔者不建议将其用于长期追踪数据的发展。

表 1 不同平滑系数取值下各类误差值的大小

Tab.1 The magnitude of errors under different smoothing coefficient

条目	平滑系数取值			
	0.1	0.2	0.3	0.4
SSE	15 533	12 069	11 568	12 268
MSE	337.7	262.4	251.5	266.7
RMSE	18.38	16.2	15.86	16.33
MAPE	47.15	41.1	39.92	41.33

2.2.2 GM(1,1)模型结果 GM(1,1)模型的预测误差高达 51.27%，后验差值 C=0.58，拟合的模型仅为基本合格。该结果的产生可能与 GM(1,1)模型适用于波动不大、经一次累加后呈指数规律且极差不应太大的数据有关。此外，GM(1,1)建模并不是数据越多越好，一般而言 4 个以上即可，且仅适合 1、2 个数据的短期预测，不适合中长期预测。

2.2.3 Markov 模型结果 Markov 模型中，根据每月痛风患者所占构成将其划分为 5 个状态，组距公式为，精确度为0.5‰，则各状态的范围为：状态 1<1‰，1‰<状态 2<3.5‰，3.5‰<状态 3<11‰，11‰<状态 4<33.5‰，状态 5>33.51‰。根据初始状态概率向量，计算得出未来 5 期处于状态 5 的概率最大，说明未来痛风均处于高患病状态。

2.2.4 ARIMA(p,d,q)模型 ARIMA(p,d,q)模型为一阶差分后平稳的非白噪声序列。且自相关系数和偏自相关系数均呈现一阶结尾特性，故模型定阶为 ARIMA(1,1,1)。但参数估计结果显示，当模型为 ARIMA(1,1,1)时参数和常数均无意义。因此可拟合模型有 ARIMA(1,1,0)和 ARIMA(0,1,1)。但 Minic 最小信息准则结果显示当模型为 ARIMA(1,1,0)时具有最小的 BIC 值。且该模型经检验后具有统计学显著性，由此确定最优模型为 ARIMA(1,1,0)。其表达式为：

$$(1-B)x_t=\frac{1}{1+0.44033B}\varepsilon_t$$

2.3 4 种模型预测效果比较

通过观察图 2 及比较表 2 的平均百分绝对误差可知，ARIMA(1,1,0)预测效果最好，指数平滑法次之，而 GM(1,1)模型预测效果最差。

表 2 不同模型预测效果比较表格

Tab.2 Comparison of prediction effect among four different models

时间	实际人数	预测值		
		ARIMA(1,1,0)模型	GM(1,1)模型	指数平滑模型
2015年1月	79	91	147	89
2015年2月	47	88	153	90
2015年3月	87	89	160	90
2015年4月	107	88	167	91
2015年5月	90	89	175	92
MAPE	-	24.72	109.21	24.95

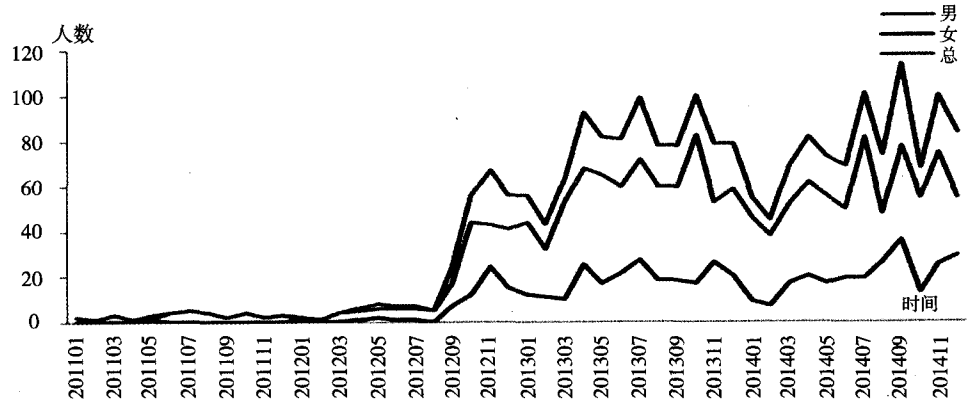


图 1 痛风月发病人数时序图

Fig.1 Timing chart of monthly gout onset number

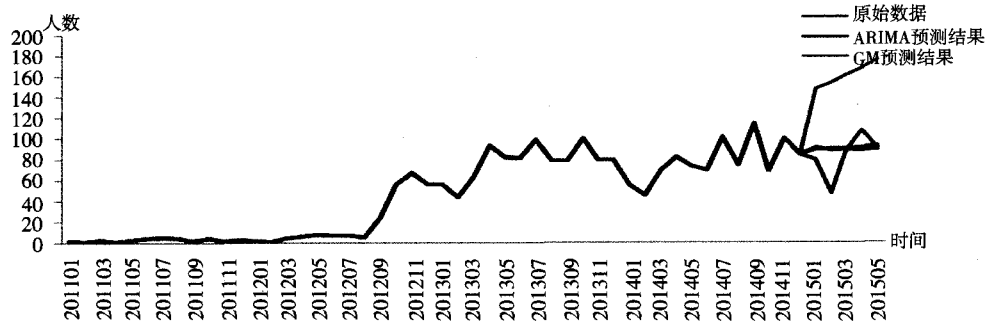


图 2 预测模型效果图

Fig.2 Effect picture of model prediction

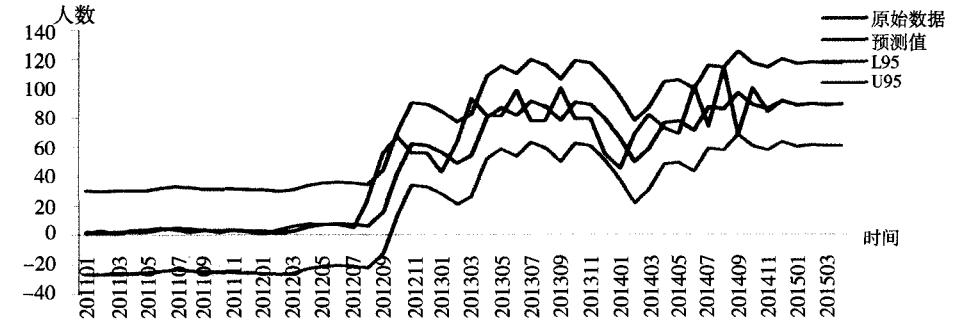


图 3 ARIMA(p,d,q)模型回代预测图

Fig.3 Back to generation prediction of ARIMA(p,d,q) model

2.4 ARIMA(1,1,0)模型回代预测结果

回代预测结果(图 3)显示：未来 5 期痛风月发病人数均为 89 人，模型平均误差为 32.15%。预测值仍在 95%的可信范围内波动，序列平稳增长。

3 讨 论

痛风是尿酸生成增多或排泄减少所致疾病，属代谢性风湿病范畴。痛风危害极大，除能导致关节病变外，高尿酸血症还与肾脏病变、尿路病变、代谢综合征、心脑血管病变、脂代谢异常、糖尿病及高血压等疾病密切相关[1,13-14]。全球痛风患病分布不均，最低不足 1%，最高已达 15.3%[15]，此外痛风患病还存在地域、性别和年龄的不均衡。总体而言，痛风患病以欧美国家较高。据第三次美国国家健康与营养调查结果显示，全美成人痛风患者 830 万，患病率为 3.9%[16]；亚洲地区既往痛风患病率较低，但随经济增长，痛风患病率也存在一定程度增加。陈玉群[17]、蒋宝泉[18]和谭琳琳[19]等报道称高尿酸血症在重庆市高收入人群、公职人员和青年招工人群中的患病率分别为 35.54%、16.01%和 13.01%。重庆市高尿酸血症患病率高，不仅与经济状况有关，还因为重庆市居民饮食习惯独特，喜爱火锅搭配啤酒。由于啤酒和火锅的主要菜品嘌呤含量均较高[19]，使得重庆市高尿酸血症高发。

既往对于痛风的研究都是基于其发病机制和治疗方案，为控制痛风的发病势头，还有必要进行趋势预测。一定时间内，痛风发病频率将组成一组随机序列，若需揭示该序列随时间的变化规律，则需进行时间序列分析。时间序列分析是定量预测方法的一种，它以概率统计论为基础来探索数据变化规律，能够对影响因素未知或资料不便收集的研究对象进行预测[20]。时间序列预测方法众多，但每种方法假设基础不同，因此有必要使用多种效能和适用条件不同的预测方法对同一批数据进行拟合，从而建立最优预测模型。

本研究使用指数平滑法、ARIMA(p,d,q)、GM(1,1)和 Markov 4 种模型对数据进行拟合。由于指数平滑法未考虑序列随机波动，且平滑系数是由研究者主观确定的，因此该方法主观性较强。此外本研究中平滑系数取值为 0.3，使得预测结果较为滞后。GM(1,1)模型适合于波动不大且经 1 次累加后呈指数规律的数据。由于本研究生成序列趋势较平，

且月发病人数极差为 113 人，综上导致 GM(1,1)模型预测效果差。马尔科夫模型用于状态和参数都离散的随机过程，其预测结果只能显示未来发病处于某个状态的概率，不能给出具体的预测值。ARIMA(1,1,0)模型综合考虑了序列的确定性信息和随机波动，能够较好地模拟数据的变化趋势，且其预测结果具有最小平均百分绝对误差，因此本研究采用 ARIMA(1,1,0)进行痛风月发病趋势预测。

ARIMA(p,d,q)模型预测精度较高，而且建模方法也比较简单，该模型使用过程中需注意以下几点[21-24]：①研究对象应满足一定的量，一般而言有 30 个以上的数据点或 7~8 个数据周期；②平稳非白噪声序列是模型使用的前提，但平稳化过程中要防止过差分现象发生；③模型识别和定阶需依据 ACF 和 PACF，如果非零的自相关系数和偏自相关系数衰减到小值波动的过程非常突然，则视为截尾；④参数估计的方法有 3 种，其中以最小二乘的精度最高，所以本研究使用最小二乘估计；⑤当存在多个模型可供选择时，应选择具有最小 AIC 或 BIC 值的模型。

本研究利用 ARIMA(p,d,q)模型进行建模获得了较好的预测效果，但时间序列 ARIMA(p,d,q)模型是利用事物的惯性发展原理进行建模，因此其短期预测精度较高。若想进行长期预测，则需不断加入新数据修正原始模型或拟合新模型，尚可获得满意效果。此外，本研究仅比较了单个预测模型的效果，仅反映了数据某一方面的特征；建立预测精度更高的组合预测模型则是本研究的后续方向，相关结果将在别文中阐述。

参 考 文 献

[1] 宋 薇，刘精东.高尿酸血症和痛风的流行病学及影响因素研究进展[J]. 江西医药，2013，48(5)：459-462.

[2] Tu FY，Lin GT，Lee SS，et al. Prevalence of gout with comorbidity aggregations in southern Taiwan[J]. Joint Bone Spine，2015，82(1)：45-51.

[3] Dalbeth N，Collis J，Gregory K，et al. Tophaceous joint disease strongly predicts hand function in patients with gout[J]. Rheumatology (Oxford)，2007，46(12)：1804-1807.

[4] Kleinman NL，Brook RA，Patel PA，et al. The impact of gout on work absence and productivity[J]. Value Health，2007，10(4)：231-237.

[5] Winnard D，Wright C，Taylor WJ，et al. National prevalence of gout derived from administrative health data in Aotearoa New Zealand[J]. Rheumatology (Oxford)，2012，51(5)：901-909.

[6] 彭彩碧. 痛风患者胰岛素抵抗、胰岛细胞功能与血尿酸水平的关系[D]. 武汉：武汉大学，2014.

[7] 向 姝，乔旭霞，罗海霞，等. 重庆市巴南区体检人群高尿酸血症患病率现状调查及与血脂相关性的研究[J]. 现代医学，2016，44(5)：649-652.

[8] 董 鹏，宋 慧. 痛风发病机制研究进展[J]. 基础医学与临床，2015，35(12)：1695-1699.

[9] 青玉凤，周京国. 痛风发病机制及药物治疗研究现状及展望[J]. 中国临床医师杂志，2012，6(10)：2719-2723.

[10] Heo SH，Lee SH. High levels of serum uric acid are associated with silent brain infarction[J]. J Neurological Sci，2010，297(1-2)：6-10.

[11] Lippi G，Montagnana M，Franchini M，et al. The paradoxical relationship between serum uric acid and cardiovascular disease[J]. Clin Chim Acta，2008，392(1-2)：1-7.

[12] Yen CJ，Chiang CK，Ho LC，et al. Hyperuricemia associated with rapid renal function decline in elderly Taiwanese subjects[J]. J Formos Med Assoc，2009，108(12)：921-928.

[13] Wijnands JM，Viechtbauer W，Thevissen K，et al. Determinants of the prevalence of gout in the general population：a systematic review and meta-regression[J]. Eur J Epidemiol，2015，30(1)：19-33.

[14] 宋 薇，刘精东，陈志雄，等. 南昌地区 40 岁以上社区居民高尿酸血症和痛风的患病率调查及相关影响因素分析[J]. 中国全科医学，2014，17(2)：181-183.

[15] 方卫纲，黄晓明，王 玉，等. 北京地区部分人群痛风的流行病学调查[J]. 基础医学与临床，2006，26(7)：781-785.

[16] Zhu Y，Pandya BJ，Choi HK. Prevalence of gout and hyperuricemia in the US general population：the National Health and Nutrition Examination Survey 2007-2008[J]. Arthritis Rheum，2011，63(10)：3136-3141.

[17] 陈玉群. 重庆市城市体检人群高尿酸血症的患病情况及相关因素分析[J]. 第三军医大学学报，2008，30(11)：1013-1017.

[18] 蒋宝泉，邓鑫钟，丕 洪，等. 重庆市某公职人员人群高尿酸血症患病情况调查研究[C]. 青岛：第五届全国中西医结合营养学术会议，2014.

[19] 谭琳琳，任 君，何华平，等. 重庆市体检人群高尿酸血症的调查分析[J]. 国际检验医学杂志，2010，31(10)：1101.

[20] 陈 勇，陈建国，朱 健，等. 江苏省启东市 1972—2001 年肺癌发病趋势分析及预测模型比较研究[J]. 中华流行病学杂志，2005，26(12)：955-959.

[21] 金如锋，黄成钢，邱 宏，等. 4 种模型对我国某地区肺结核发病率的预测[J]. 现代预防医学，2008，35(24)：4866-4869.

[22] 张彦琦，唐贵立，王文昌，等. ARIMA 模型及其在肺结核预测中的应用[J]. 现代预防医学，2008，35(9)：1608-1610.

[23] 吴家兵，叶临湘，尤尔科. ARIMA 模型在传染病发病率预测中的应用[J]. 数理医药学杂志，2007，20(1)：90-92.

[24] 樊雯婧，陆 群，邹立巍，等. ARIMA 模型在合肥市疟疾发病预测中的应用[J]. 安徽医科大学学报，2013，48(3)：252-256.

（责任编辑：张辉洁）