状态空间模型在我国肝炎发病率预测中的应用

阮 俊[1],王 玥[2],郁婷燕[2],戚凯莉[3],陈冠军[1],丁 勇[4],吴 静[4*]

([1]南京医科大学生物医学工程系,[2]第一临床医学院,[3]公共卫生学院,[4]数学与计算机教研室,江苏 南京 210029)

[摘 要] **目的**:应用状态空间模型对我国甲肝、乙肝、丙肝的月发病率进行拟合和预测,分析预测准确率与发病周期规律的关系,为肝炎发病率的预测提供新方法。**方法**:对 2005—2013 年我国甲肝、乙肝、丙肝的月发病率建立状态空间模型,应用 MATLAB 软件中的 SSM 工具包对模型进行拟合,并对 2014 年 1—12 月的发病率进行预测;通过数据标准化,建立发病率周期规律性评价指标。**结果**:甲肝、乙肝和丙肝的月发病率预测相对误差的平均值分别为 17.03%、6.30%和 2.03%,标准化数据周期标准差均值分别为 0.214 3、0.195 2 和 0.172 4。**结论**:状态空间模型能较好地应用于肝炎发病率的拟合和预测,标准化数据周期评价指标的标准差越小,预测效果越好。

[关键词] 状态空间模型;肝炎;预测;周期性

[中图分类号] R512.6 [文献标志码] A [文章编号] 1007-4368(2016)03-380-05

doi:10.7655/NYDXBNS20160327

Application of state space model in forecasting the incidence of hepatitis in China

Ruan Jun[1], Wang Yue[2], Yu Tingyan[2], Qi Kaili[3], Chen Guanjun[1], Ding Yong[4], Wu Jing[4*]

([1]*Department of Biomedical Engineering*, [2]*the First Clinical Medical College*, [3]*School of Public Health*, [4]*Department of Mathematics and Computer*, *NJMU*, *Nanjing 210029*, *China*)

[**Abstract**] **Objective**: To forecast the monthly incidence of hepatitis A, B and C in China by state space model. To analyze the relationship between the accuracy of prediction and the cycle regularity of the incidence, and to provide a new approach to forecast the incidence of hepatitis. **Methods**: State space model was fitted with data of monthly reported cases of hepatitis A, B and C in China from 2005 to 2013. The SSM toolbox of MATLAB software was performed to construct the state space model, and the constructed model was applied to predict the monthly incidence of 2014. The evaluation index of the cycle regularity of the incidence was established by standardizing data. **Results**: The average relative error of prediction of hepatitis A, B and C was 17.03%, 6.30% and 2.03%, respectively, and the mean periodic standard deviation of the standard data was 0.2143, 0.1952 and 0.1724, respectively. **Conclusion**: The state space model can be fitted and performed to make a short-term prediction of the incidence of hepatitis. When the periodic standard deviation of the standardized data is smaller, the predicted result is more accurate.

[**Key words**] state space model; hepatitis; prediction; periodicity

[Acta Univ Med Nanjing, 2016, 36(03): 380-384]

肝炎最常见的原因是病毒感染。病毒性肝炎分为甲、乙、丙、丁和戊型,虽然病毒种类不同,但都足以对人构成严重危害,其中乙型和丙型肝炎可以导致肝硬化和肝癌的发生。全世界每年约有一百万人死于与病毒性肝炎相关的疾病,其中最常见的是肝硬化和肝癌。目前,全世界数百万人患有病毒性肝炎,更多人有被感染的风险[1]。据统计,我国是全球肝硬化发病率、死亡率最高的国家。肝炎病毒感染者的预后,取决于病程长短及治疗结果,越早就医、并且治疗彻底的患者病程越短,预后越好。中国肝炎防治基金会提出了“早治肝炎,告别肝炎”的口号。因此,探讨肝炎的流行规律、预测其发病率对肝炎的防治工作有着重要的指导意义[2]。

近年来,状态空间模型作为一种新的建模工具,被用来估计不同的状态成分以达到分析和预测

[基金项目] 江苏省大学生实践创新训练计划项目(201310312036Y);江苏省高校自然科学基金(13KJB310007);南京医科大学科技发展基金重点项目(2013NJMU006)

*通信作者(Corresponding author),E-mail:wujing@njmu.edu.cn

的目的,已应用于经济、工程、生物等多个领域[3-9],在医学领域对传染病进行预测研究也已经开始[10]。本文尝试用该模型对我国甲肝、乙肝和丙肝发病率同时进行建模,预测其未来的发病趋势,提出评价传染病数据周期性规律的指标,分析发病周期与预测准确性的关系,为早期发现肝炎及制定相关的防治策略提供依据。

1　材料和方法

1.1　材料

甲肝、乙肝和丙肝的传染病数据资料来源于我国卫计委疾病预防控制局(http://www.nhfpc.gov.cn/jkj/s2907/list.shtml)2005 年 1 月—2014 年 12 月的全国法定报告传染病疫情资料，其中 2005 年 1 月—2013 年 12 月的数据用于建立模型,2014 年 1—12 月数据用于验证模型的预测效果。全国人口数据资料来源于国家统计局网站 http://wwww.stats.gov.cn。本研究用月发病率数据(单位:1/10 万)进行分析。

1.2　方法

1.2.1　建模方法

能够完全表征系统动力学特征的一组独立变量称为系统的状态变量，它是系统的内部变量,在许多实际问题中,这些状态变量一般都是无法观测到的变量,反映了系统所具有的真实状态。状态向量取值的空间(即以状态向量为坐标构成的空间)称为状态空间。状态空间模型建立了可观测量和系统内部变量之间的关系,从而可以通过估计各种不同的状态向量达到分析和预测的目的。

一般的线性高斯状态空间模型由两个方程构成[3]:

状态方程 $x_t=B_t x_{t-1}+\delta_t$

量测方程 $y_t=A_t x_t+\varepsilon_t$　　(1)

$t=1,2,\cdots,n$

其中 t 表示时间,y_t 是可观测量,本文表示不同时间的传染病发病率。模型假定 y_t 由趋势项、循环项、季节项和不规则项构成;x_t 为 $m\times1$ 维状态向量;B_t 为 $m\times m$ 阶状态转移矩阵;A_t 为 $1\times m$ 阶量测矩阵;ε_t 为量测方程的随机扰动项;δ_t 为状态方程的随机扰动向量。

求解状态空间模型的核心算法是 Kalman 滤波[3,10]。Kalman 滤波的主要作用是,当随机误差项和初始状态向量服从正态分布时,能通过预测误差分解计算似然函数,从而达到对模型中所有未知参数进行估计的目的。

本文按文献[10]建立和拟合状态空间模型:$m=13$,$\delta_t=[\zeta_{1t},\zeta_{2t},\omega_t,0,\cdots,0]'$,$\varepsilon_t\sim N(0,\sigma_\varepsilon^2)$,$\zeta_{1t}\sim N(0,\sigma_{\zeta1}^2)$,$\zeta_{2t}\sim N(0,\sigma_{\zeta2}^2)$,$\omega_t\sim(0,\sigma_\omega^2)$。随机扰动项的方差 σ_ε^2、$\sigma_{\zeta1}^2$、$\sigma_{\zeta2}^2$、σ_ω^2 较小,说明模型拟合较好;采用 MATLAB 软件中工具箱(SSM 程序包)进行数据的处理和分析[10-11]。

1.2.2　数据的标准化处理

为了考察传染病数据的周期性,需要对数据进行标准化处理。

设原始数据为 $x_{ij}>0$,i 代表年,j 代表月;$i=1,2,\cdots,k$;$j=1,2,\cdots,12$。对确定的 i,记$\{x_{ij}\}$的最大值为 x_{imax},最小值为 x_{imin}。数据标准化处理采用公式:

$$x'_{ij}=\frac{x_{ij}-x_{imin}}{x_{imax}-x_{imin}}\qquad(2)$$

经过标准化处理，所有数据在 0 和 1 之间,从而具有可比性。

上述方法为线性变换,具有良好性质:保持数据变换前后的大小顺序不变;保持数据变换前后的间距之比为常数。

2　结　果

2.1　模型拟合

用 matlab 软件将甲肝、乙肝和丙肝数据进行状态空间模型拟合和计算,得到参数见表 1。

表 1　模型参数

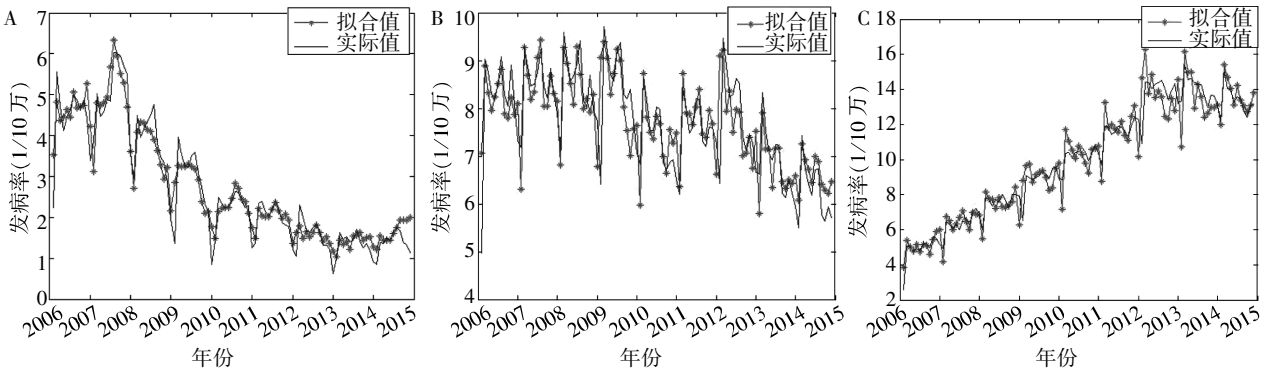
Table 1　Model　parameter

参数名称	甲肝	乙肝	丙肝
σ_ε^2	1.096×10^{-2}	2.068×10^{-1}	5.078×10^{-1}
$\sigma_{\zeta1}^2$	6.275×10^{-2}	1.325×10^{-2}	1.984×10^{-3}
$\sigma_{\zeta2}^2$	1.160×10^{-11}	1.528×10^{-4}	3.260×10^{-4}
σ_ω^2	5.449×10^{-3}	1.144×10^{-9}	4.351×10^{-3}

各 $\sigma^2\approx0$,说明状态空间模型是可行的,模型拟合效果见图 1,其中蓝色为原始数据,红色为拟合数据,横轴为实际年份,纵轴为发病率。

2.2　模型的预测

用状态空间模型预测我国 2014 年 1—12 月甲肝、乙肝和丙肝逐月发病率,并结合卫生部网站公布的实际数据进行预测精度的验证(表 2)。进一步对表 2 中的 3 列相对误差作了方差分析,结果显示 $P=5.06\times10^{-4}<0.05$,有显著差异,说明用状态空间模型分别对甲肝、乙肝、丙肝的月发病率进行预测时,产生的相对误差不一样。



A:甲肝;B:乙肝;C:丙肝。

图 1 状态空间模型的发病率拟合图

Figure 1 Fitted value series of incidence by state space model

表 2 2014 年月发病率的实际值与预测值

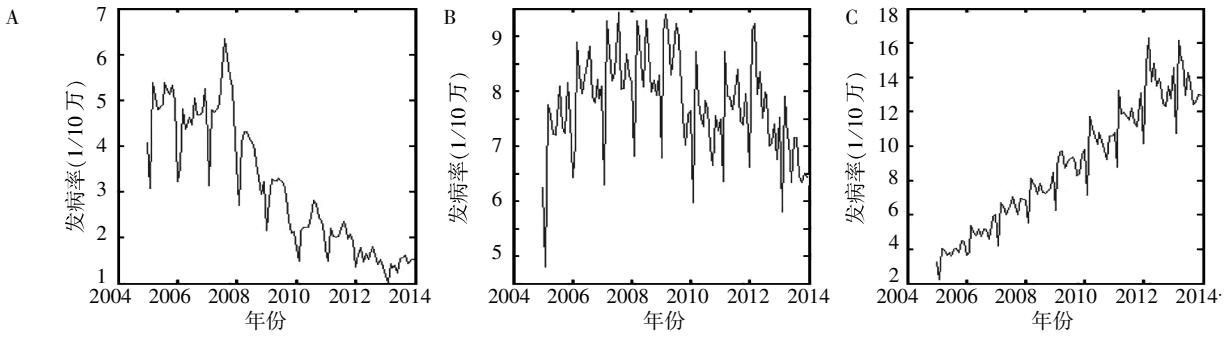
Table 2 Actual values and predictive values of monthly incidence in 2014 (1/10 万)

	甲肝			乙肝			丙肝		
月份	实际值	预测值	相对误差(%)	实际值	预测值	相对误差(%)	实际值	预测值	相对误差(%)
1	1.274 3	0.919 7	27.83	6.595 2	5.960 3	9.63	13.108 4	12.926 4	0.38
2	1.222 4	0.840 0	31.27	6.073 0	5.492 1	9.57	12.006 7	12.204 8	1.65
3	1.544 8	1.374 4	11.04	7.259 1	7.426 3	2.30	15.399 7	15.399 7	1.75
4	1.444 6	1.362 8	5.66	6.928 4	6.724 1	2.95	14.704 4	14.055 2	4.41
5	1.440 2	1.506 0	4.59	6.721 4	6.416 0	4.54	14.023 8	13.939 7	0.60
6	1.444 6	1.410 3	2.36	6.448 3	6.186 5	4.06	13.126 7	13.278 1	1.15
7	1.602 6	1.672 7	4.40	6.992 7	6.712 9	4.00	14.211 7	13.601 2	4.29
8	1.759 7	1.756 8	0.17	6.877 7	6.726 8	2.20	13.414 8	13.436 0	0.16
9	1.941 8	1.665 4	14.24	6.420 9	5.788 6	9.85	13.060 9	12.746 6	2.41
10	1.941 8	1.372 3	29.32	6.287 1	5.634 7	10.38	12.747 3	12.399 3	0.08
11	1.934 5	1.327 7	31.38	6.223 4	5.941 6	4.53	13.071 2	13.082 1	4.70
12	1.998 1	1.142 7	42.81	6.462 6	5.708 7	11.67	13.812 5	13.164 0	1.70
均值			17.09			6.30			2.03

由表 2 可知,模型的预测结果,甲肝相对误差最大为 42.81%,最小为 2.36%,平均为 17.09%;乙肝相对误差最大为 11.67%,最小为 2.20%,平均为 6.30%;丙肝相对误差最大为 4.70%,最小为 0.08%,平均为 2.03%。可见状态空间模型能较好地对肝炎发病率进行预测,预测较好的是乙肝和丙肝。

2.3 季节性影响

图 2 是 2005 年 1 月—2013 年 12 月我国甲肝、乙肝和丙肝月发病率时间序列图,从图中可以看出,甲肝发病率总体呈下降趋势,一般在每年的 3 月前后常会出现发病率高峰;乙肝发病率总体呈下降趋势,在每年的 3 月份和 8 月份常会出现发病



A:甲肝;B:乙肝;C:丙肝。

图 2 2005—2013 年发病率时序图

Figure 2 Time series of monthly incidence of hepatitis from 2005 to 2013 in China

率高峰;丙肝发病率总体呈上升趋势,一般在每年的 3 月前后常会出现发病率高峰。因此,肝炎的发病率有两个特征:一是上升和下降的变化趋势;二是以年为单位的季节性变化周期性。

从数学角度来看,周期函数经过一个周期之后,函数值大小不变,$f(x)=f(x+T)$,T 为最小周期。而我们看到的传染病发病率具有的周期性,比一般的周期性更复杂,还伴随着数值上升和下降的变化。为了评价这种周期性规律,我们提出如下的评价指标:首先对数据进行标准化处理,以消除数据上升和下降的影响,使在不同周期时间段的数据具有可比性;再根据一般周期函数的定义,如果数据具有较好的周期性,标准化数据在每个周期的同一时间点(月)应该相同,或者它们的差别应尽可能小,这个差别我们用标准差来评价,记 s_j 为第 j 月标准化数据的标准差,则平均标准差为 $s=\frac{\sum_{j=1}^{12} s_j}{12}$,$s$ 越小,说明数据的周期性越好。

用公式(2)计算甲肝、乙肝和丙肝标准化处理后数据每月的标准差见表 3。

乙肝、丙肝的周期性要优于甲肝。以表 3 中的标准差均值为横坐标,表 2 中的相对误差均值为纵坐标作散点图(图 3)。显示状态空间模型的预测效果与样本数据的周期性动态变化趋势有关,周期性变化趋势越一致,预测结果也越准确。

表 3　标准化处理后数据每月的标准差

Table 3　Periodic standard deviation of the standardized data monthly

月份	甲肝	乙肝	丙肝
1	0.195 3	0.306 7	0.288 2
2	0.267 4	0.395 8	0.320 5
3	0.189 2	0.040 3	0.092 3
4	0.238 6	0.120 1	0.127 7
5	0.172 6	0.053 1	0.090 0
6	0.221 7	0.173 4	0.122 0
7	0.118 4	0.153 7	0.135 4
8	0.088 2	0.179 1	0.155 9
9	0.155 6	0.174 1	0.178 5
10	0.255 3	0.189 7	0.128 1
11	0.321 1	0.279 6	0.190 1
12	0.348 3	0.277 3	0.240 4
均值	0.214 3	0.195 2	0.172 4

3　讨　论

目前,在我国各级疾病预防控制中心的工作人员和医学院校的科研人员从不同疾病、不同层次、不同方法对传染病的预测预警进行了大量研究,所用方法多种多样。现有的定性预测方法有控制图法、比数图法、模糊数学理论、马尔可夫链预测法等。现有的定量预测方法有灰色动态模型、回归预测模型、微分方程模型、灰色预测模型、余弦模型等。

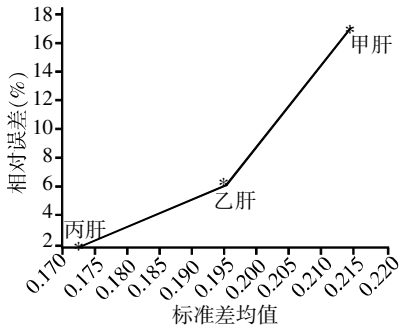


图 3　甲肝、乙肝、丙肝的标准差均值与预测相对误差的关系图

Figure 3　Relationship between mean standard deviation and relative error of prediction of hepatitis A, B and C

传染病发生往往具有长期趋势、季节性、周期性、短期波动和不规则变动等特征,是个复杂的时间序列,因此对这类资料进行预测分析时需要同时考虑这些特征。而目前应用这些模型无法全面考虑这些特征,导致对复杂时间序列进行分析的准确度下降。状态空间模型是一种新型预测方法,其基本思想是利用时间序列的观测值所具有的依存关系或自相关性,把模型更加细节化,通过趋势项、循环项、季节项和不规则项等因素能更有效地分析与预测,因此较适合于传染病的分析与预测。本研究对甲肝、乙肝和丙肝的应用结果,说明了状态模型的可行性。比较于林凤等[12]用 ARIMA 季节模型对我国丙肝发病进行的预测,状态空间模型预测结果更加准确(其预测的相对误差最大为 17.60%,平均为 7.80%;本研究最大为 4.70%,平均为 2.03%)。陈友春等[10]用状态空间模型对我国肺结核发病率进行预测,相对误差小于 10%,也说明了该模型的可行性。

应用数学模型进行传染病研究,预测的准确性是一个重要问题,本研究分析了数据周期规律性和预测准确性的关系。传染病的发病往往与季节有关,具有较明显的周期性[13-15]。状态空间模型考虑了周期性因素,因此发病规律的周期性动态变化对模型预测效果有影响。显然,周期性变化越有规律性,预测效果也会更好。甲型肝炎具有爆发流行的特点,因此发病的周期性相对于乙肝和丙肝要差一些,所以

预测效果也差一些，这给疾病预测带来一定难度。比较表 2 和表 3 的结果，也验证了这一点。

传染病数据除了周期性变化之外，还有总体上升或下降趋势，数据分析比较复杂。本研究利用数据标准化变换，消除了数据上升或下降的趋势，使数据的周期性可以进行比较，在此基础上，提出了评价周期性的指标。是否还有更好能用于传染病数据周期分析的评价方法，值得进一步探讨。

传染病预测预警工作的研究方法和理论已经取得了长足进步，逐渐走向成熟。状态空间模型这一新方法在传染病方面的应用，将完善和充实目前的传染病监测体系，提高今后传染病预防控制工作的预见性和主动性，从而为更好地维护人民群众的身体健康，提供了一种可行方法。由于新方法的应用还处于起步阶段，因此还需对理论和计算、模型应用的特点、模型参数的选取，提高模型的拟合和预测效果等各方面工作进行完善。

[参考文献]

[1] El-Serag HB. Epidemiology of viral hepatitis and hepatocellular carcinoma[J]. Gastroenterology,2012,142 (6):1264-1273.e1

[2] 张敏娜，袁 月，貌盼勇，等. 中国 2004—2013 年病毒性肝炎发病与死亡趋势分析[J]. 中华流行病学杂志，2015,36(2):144-147

[3] Kalman RE. A new approach to linear filtering and prediction problem[J]. J Basic Eng,1960,82(1):35-45

[4] Durbin J,Koopman SJ,Time series analysis by state space methods[M]. New York:Oxford University Press,2001:23

[5] 仇伟杰. 基于状态空间模型与 KALMAN 滤波的中国电力需求分析[J]. 工业技术经济，2006,25(2):63-66

[6] Patterson TA,Thomas L,Wilcox C,et al. State-space models of individual animal movement[J]. Trends Ecol Evol,2008,23(2):87-94

[7] 高铁梅，王金明，陈 飞. 中国转轨时期经济增长周期波动特征的实证分析[J]. 财经问题研究，2009(1):22-29

[8] 李 勇，王有贵. 基于状态空间模型的中国房价变动的影响因素研究[J]. 南方经济，2011(2):38-45

[9] 陈真玲，鲁丰先，王光辉. 基于状态空间模型的我国行政管理费动态分析[J]. 北京理工大学学报(社会科学版)，2014,16(6):78-84

[10] 陈友春，朱文婕. 状态空间模型及其在传染病发病率预测中的应用[J]. 南京医科大学学报(自然科学版)，2015,35(2):275-278

[11] Peng JY,Aston JD. The state space models toolbox for MATLAB[J]. J Stat Softw,2011,41(6):1-26

[12] 于林凤，吴 静，周锁兰，等. ARIMA 季节模型在我国丙肝发病预测中的应用[J]. 郑州大学学报(医学版)，2014,49(3):344-348

[13] 王 超，丁 勇，陆 群，等. ARIMA 乘积季节模型在我国甲肝发病预测中的应用[J]. 南京医科大学学报(自然科学版)，2014,34(1):75-79

[14] 朱 平，马 平，张 烽，等. 南通市 2004~2012 年戊型病毒性肝炎流行特征分析[J]. 南京医科大学学报(自然科学版)，2014,34(3):385-387

[15] 胡建利，祖荣强，彭志行，等. 江苏省戊型肝炎发病趋势的时间序列模型应用[J]. 南京医科大学学报(自然科学版)，2011,31(12):1874-1878

[收稿日期] 2015-09-07

(上接第 367 页)

[3] Chestovich PJ,Lin AY,Yoo J. Fast-track pathways in colorectal surgery[J]. Surg Clin North Am,2013,93 (1):21-32

[4] Kehlet H. Fast-track surgery-an update on physiological care principles to enhance recovery[J]. Langenbecks Arch Surg,2011,396(5):585-590

[5] Salhiyyah K, Elsobky S, Raja S, et al. A clinical and economic evaluation of fast-track recovery after cardiac surgery[J]. Heart Surg Forum,2011,14(6):E330-E334

[6] Reif P, Drobnitsch T, Aigmueller T, et al. The decreasing length of hospital stay following vaginal hysterectomy: 2011-2012 *vs.* 1996-1997 *vs.* 1995-1996 [J]. Geburtshilfe Frauenheilkd,2014,74(5):449-453

[7] Carter J. Fast-track surgery in gynaecology and gynaecologic oncology:a review of a rolling clinical audit[J/OL]. ISRN Surg,2012,2012:368014 [2016-01-19]. http://www.ncbi,nlm,nib,gov/pmc/articles/PMC3540771. DOI:10.5402/2012/368014

[8] Kjolhede P, Langstrom P, Nilsson P, et al. The impact of quality of sleep on recovery from fast-track abdominal hysterectomy[J]. J Clin Sleep Med,2012,8(4):395-402

[9] 余继海，许戈良，马金良，等. 损伤控制和加速康外科理念在原发性肝癌合并肝硬化手术治疗中的价值[J]. 肝胆外科杂志，2010,18(1):19-22

[10] 范 烨，壮 麟，钱晓峰，等. 加速康复外科治疗在肝移植中的应用价值[J]. 器官移植杂志，2014,5(6):349-351

[11] 饶建华，吕 凌，王 平，等. 腹腔引流术在肝脏切除术后应用的必要性探讨[J]. 中华普通外科杂志，2010,25(4):303-305

[收稿日期] 2015-11-05