JRHS
Journal of Research in Health Sciences
journal homepage: www.umsha.ac.ir/jrhs

Original article

Predicting Hepatitis B Monthly Incidence Rates Using Weighted Markov Chains and Time Series Methods

Maryam Shahdoust (MSc)[a], Majid Sadeghifar (PhD)[b], Jalal Poorolajal (MD, PhD)[c], Niloofar Javanrooh (MSc)[a], and Payam Amini (MSc)[a*]

[a] *Department of Biostatistics & Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*

[b] *Department of Statistics, Faculty of Basic Science, Bu-Ali Sina University, Hamadan, Iran*

[c] *Chronic Diseases (Home care) Research Center and Department of Epidemiology & Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*

ARTICLE INFORMATION

Article history:
Received: *14 July 2014*
Revised: *18 October 2014*
Accepted: *07 December 2014*
Available online: *09 December 2014*

Keywords:
Hepatitis B
Markov Chains
Time series
Incidence rate
Prediction

**** Correspondence***
Payam Amini (MSc)
Tel: *+98 919 7552131*
E-mail: *p.amini@umsha.ac.ir*

ABSTRACT

Background: Hepatitis B (HB) is a major global mortality. Accurately predicting the trend of the disease can provide an appropriate view to make health policy disease prevention. This paper aimed to apply three different methods predict monthly incidence rates of HB.

Methods: This historical cohort study was conducted on the HB incidence data of Hamadan Province, the west of Iran, from 2004 to 2012. Weighted Markov Chain (WMC) method based on Markov chain theory and two time series models including Holt Exponential Smoothing (HES) and SARIMA were applied on the data. The results of different applied methods were compared to correct percentages of predicted incidence rates.

Results: The monthly incidence rates were clustered into two clusters as state of Markov chain. The correct predicted percentage of the first and second clusters for WMC, HES and SARIMA methods was (100, 0), (84, 67) and (79, 47) respectively.

Conclusions: The overall incidence rate of HBV is estimated to decrease over time. The comparison of results of the three models indicated that in respect to existing seasonality trend and non-stationarity, the HES had the most accurate prediction of the incidence rates.

Citation: Shahdoust M, Sadeghifar M, Poorolajal J, Javanrooh N, Amini P. Predicting Hepatitis B Monthly Incidence Rates Using Weighted Markov Chains and Time Series Methods. J Res Health Sci. 2015; 15(1): 28-31.

Introduction

Hepatitis B (HB) is one of the prevalent diseases in world and a major cause of morbidity and mortality[1]. More than two billion people alive today have been infected with hepatitis B virus (HBV). Among Asian countries, Iran is considered as a country with a low endemicity. According to studies conducted in 2009, the prevalence of chronic HBV in general population of Iran was estimated about 1.7% [2]. Furthermore, due to mass vaccination since 1993, prevalence of chronic hepatitis B infection has being decreased considerably. For example, in Hamadan Province, the west of Iran, the incidence rate was decreasing during 2004 to 2009 from 19.6 to 7.7 (per 100,000 populations) [3]. The high prevalence of infection could represent a failure of public health considering the availability and effectiveness of HBV immunization for all generation. Therefore, accurately forecasting the incidence rate could provide a measure to evaluate the public health programs and help to prevent and control the disease.

Statistical methods are frequently used methods in epidemiological studies. The most frequently used methods in studying infectious disease such as HBV are regression models[4,5] time series[6,7] multivariate analysis[8, 9] and also machine learning methods[10,11].

Among statistical methods Markov chain process have appropriate properties to study the dynamic behavior of diseases[12-14]. Markov chains are an important class of stochastic processes in which a future state of an experiment depends only on the present one, not on proceeding states[13,14]. Markov chains and Markov processes are extensively applied in medical researches[13,15] such as prediction of AIDs epidemic[16].

In 2010, Chen et al. proposed an approach according to weighted Markov chain to forecast and analysis the incidence of infectious diseases[17]. The proposed approach applies the sequential cluster method at incidence rates to determine the state space of Markov chain. Then, future incidence states would be predicted by using weighted Markov chain. In this approach, weights are standardized self-correlation coefficients of different orders. Their method successfully has been validated by existing incidents data of HBV in Jiangsu Province. They indicated that the proposed approach

could overcome the difficulties of other methods such as dependency on historical data with a high accuracy [17-19].

This study aimed to apply the Chen et al. [17] approach to predict monthly HBV incidence rates approach as a new method of prediction and compare the results with traditional methods such as HES and SARIMA models for predicting the incidence rates of HBV incidence in Hamadan Province, western Iran from 2004 to 2012.

Methods

We used the dataset of a historical cohort study, conducted on the HBV incidence in Hamadan Province, the west of Iran, from 2004 to 2012. The data were extracted from the database of the Vice-chancellor of Health Services, Hamadan University of Medical Sciences [3], Hamadan, Iran. In order to predict monthly incidence rates of HBV, two different approaches were applied including weighted Markov chain and time series models. The used time series models were Holts Exponential Smoothing and Seasonal ARIMA methods.

Weighted Markov Chain Approach (WMC)

This method, which was first proposed by Chen et al.[17] is based on traditional Markov chain theory and the correlation analysis approach. This method includes the following steps.

a) Clustering monthly incidence rates: The monthly incidence rates of HBV were partitioned into homogenous clusters using one-dimensional sequential cluster methods[17,20]. Then, the resulted clusters identify the Markov chain states so that the incidence rate of each month belongs to a certain state.

b) Calculation of self-correlation coefficients r_k:

$$r_k = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n-k}(x_i - \bar{x})^2} \qquad (2.1)$$

Where r_k is the k^{th} month self-correlation coefficient, x_i ($i= 1,\ldots,n$) is the i^{th} monthly incidence rate, $\bar{x}$ and n are mean value of x_i and the length of monthly incidence series, respectively. The weights are calculated by equation (2.1) where m is the maximum step according to prediction. i.e., $k=1,\ldots,m$.

$$w_k = | r_k | / \sum_{k=1}^{m} | r_k | \qquad (2.2)$$

c) Computing transition probabilities matrices:

The number of transition probability matrices is due to the number of determined clusters at the first step. The transition probability matrices, $P^{(i)}$ $i=1,\ldots,k$, where i is the number of transition steps and k is the number of clusters, are computed through the frequency of transitions from state to state. For example, the first array of the one-step transition matrix $P^{(1)}$ is the frequency of transitions from state one to state one in one step.

d) Calculating the predicting probabilities:

States predicting probabilities are computed using equation (2.3) where m indicates the number of months to forecast.

$$p_i = \sum_{k=1}^{m} w_k P_i^{(k)} \qquad (2.3)$$

e) Predicting monthly incidence rates state:

The predicted state of each month is determined according to the maximum of P_i's calculated at previous step.

Time series models

A time series model is a stochastic process or a set of observations y_t, that each observation is recorded at a specific time t [21]. To inference such series, it is necessary to set up a hypothetical probability model to present the data and choose an appropriate model. Then, it is possible to estimate parameters, check for goodness of fit and understand the mechanism from which the series has been come [22].

In respect of existing seasonal and stationary time series, different models could be conducted. In this paper, two different time series models, HES and Seasonal Auto Regressive Integrated Moving Average (SARIMA) were applied.

An HES is a flexible time series method for short-term forecasts[23]. It updates the level and trend parameters in non-seasonal series while assuming no assumption for error terms[24-26].

A SARIMA is classified as an ARIMA(p,d,q)(P,D,Q) model, where p, d and q are non-negative integer that refer to the order of the autoregressive, integrated and moving average parts of the ARIMA model respectively. In addition, P is the number of seasonal autoregressive term, D is the number of seasonal differences and Q indicates seasonal moving average term[22, 25]. Selecting the best SARIMA model was carried out using Akaike Information Criteria (AIC). Stationarity assumption was checked using Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test[22]. The statistical significant level was assumed as 0.05.

Based on the results of study conducted by Chen et al [17], the results of WMC approach was evaluated by computing the correct predicted probability. In order to make the three methods comparable, we used the same evaluation method for time series models.

All the analyses were done by R.3.1.0. software.

Results

The dataset included the HBV incidences from 2004 to 2012. During the study period, 1992 subjects with HBV infection were diagnosed. The mean (SD) age of the patients was 40.51 (0.36) with ranged from one to 100 yr. A majority of patients (74.11%) was male and 73.35% were married.

To predict monthly incidence rate of HBV infection during 2004-2012, weighted Markov chains method and time series models were applied.

Weighted Markov chains method results

According to the results of K-means approach, the standardized incidence rates were categorized into two groups. In order to be sure of the appropriateness of number of clusters, we used a hierarchical method entitled WARD s.

According to the results of this method, selecting two clusters was the best model. The first cluster included incidence rates from zero to 0.497 and the second cluster included the remaining higher incidence rates (0.501- 1).

According to the number of determined clusters, WMC approach could be performed using two steps. The first and second order autocorrelation measures and corresponding weights were 0.20, 0.02 and 0.88, 0.12, respectively. Various transition probabilities according to the number of steps were constructed as follows.

$$P^{(1)}=\begin{bmatrix}0.641 & 0.359\\ 0.600 & 0.400\end{bmatrix} \qquad P^{(2)}=\begin{bmatrix}0.666 & 0.334\\ 0.575 & 0.425\end{bmatrix}$$

The probability of correct incidence predictions of WMC approach is given in Table 1.

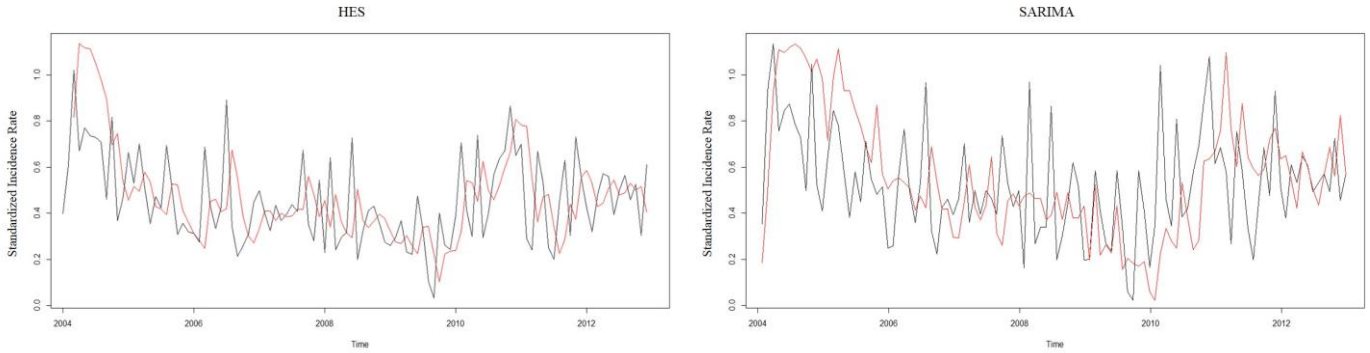
HES and SARIMA models results

After checking the seasonality and stationary assumptions, two time series models, HES and SARIMA (0,1,1) (1,0, 0) were fitted on the data. The best SARIMA model was carried out using Akaike Information Criteria (AIC). To perform HES method, the time series seasonality trend was removed by subtracting the seasonal component from observations. The estimates showed that the prediction of the current level and the slope of the trend were not highly affected by the recent observations. Figure 1 shows the smoothed plots and model equations of obtained time series, where *k* indicates number of steps. Accordingly, the overall incidence rate of HBV is estimated to decrease over time.

The correct incidence prediction probabilities of three methods are shown in Table 1.

Table 1: Correct prediction probabilities for standardized incidence rates

Methods	Correct prediction probability	
	State 1	State 2
Weighted Markov Chain(WMC)	1.00	0.00
Holt Exponential Smoothing (HES)	0.84	0.67
Seasonal Auto Regressive Integrated Moving Average (SARIMA)	0.79	0.47



$$y_{t+1}=a_t+kb_t$$
$$a_t=0.42y_t+(1-0.42)(a_{t-1}+b_{t-1})$$
$$b_t=0.26(a_t-a_{t-1})+(1-0.26)(b_{t-1})$$

SARIMA(0,1,1)(1,0,0)

Figure 1: Smoothed plots and model equations of Holt Exponential Smoothing (HES) and Seasonal Auto Regressive Integrated Moving Average (ARIMA (0,1,1) (1,0,0))- The red lines: fitted plots, The black lines: Observed plots)

Discussion

Among different statistical methods used in prediction of incidence rate, Markov chain models are rarely used. This study tried to apply WMC to predict the monthly incidence rates. Furthermore, different time series models according to the existing stationary and seasonality assumptions were fitted.

Before applying the methods, incidence rates were classified into two clusters. Most of the incidences were classified in cluster one which includes the lower incidences. In WMC method, the transitions enounced that the probability of state one to the same state transitions in one and two steps were 0.66 and 0.64, respectively. Therefore, it was expected that the monthly-predicted incidence rates would belong mostly to the first cluster rather than the second one. The WMC results showed that all incidence rates were predicted in cluster one. To study these results the homogeneity of chain were checked.

The trend of incidence rates illustrated a decreasing trend from 2004 to 2009 and slight increase from 2010 to 2012. In addition, KPSS test resulted in non-stationary of the chain (P<0.05). Therefore, the time series models, which consider the non-stationary chains, were applied. In respect to existing seasonality trend in the time series, a SARIMA model was applied. The correct prediction percentage for clusters one and two according to the model were 0.79 and 0.47, respectively. After removing seasonality from the dataset, the HES model was performed. Compared to SARIMA results, an outperformance was found by applying HES where 84% and 67% of incidence rates were correctly predicted in the first and second clusters, respectively.

Based on the correct predicted probabilities resulted for the three methods, the HES model had the greatest correct predicted probabilities followed by SARIMA and then WMC.

The main advantage of WMC is that it is less dependent on historical data, thanks to scientific classification determining the initial states and transition probabilities. Although WMC is an appropriate dynamic model for incidence predictions of infectious diseases, its efficiency strongly depends on stationarity of the time series[17, 27]. In contrast, time series model predictions strongly depend on historical data [17]. HES, as well as SARIMA allows all the observations to be contributed in prediction process.

The main limitation of this study was the short duration of follow-up time whereas, Markov chain approach has a good performance when the duration of time is long enough. In

short duration of follow up, the increase and decline in the historical data will not be fully reflected.

Conclusions

The overall incidence rate of HBV is estimated to decrease over time. The comparison of results of the three models indicated that in respect to existing seasonality trend and non-stationarity, the HES had the most accurate prediction of the incidence rates.

Acknowledgements

We thank Feng Chen and Zhihang Peng for their valuable help. We also thank the Vice Chancellor of Health Services of Hamadan University of Medical Sciences for giving us the dataset. In addition, we appreciate the Vice Chancellor of Research and Technology of Hamadan University of Medical Sciences for financial support of this study.

Conflict of interest statement

The authors have no conflict of interest to declare.

References

1. Carey WD. The prevalence and natural history of hepatitis B in the 21st century. *Cleve Clin J Med.* 2009;76(Suppl 3):S2-S5.

2. Poorolajal J, Majdzadeh R. Prevalence of chronic hepatitis B infection in Iran: a review article. *J Res Med Sci*. 2009;14(4):249-258.

3. Poorolajal J, Mirzaei M, Bathaei SJ, Majzoobi MM. Hepatitis B and C infections in hamadan province during 2004-2009. *J Res Health Sci.* 2011;11(1):51-57.

4. He WP, Hu JH, Zhao J, Tong JJ, Ding JB, Lin F, Wang HF. Comparison of four prognostic models and a new Logistic regression model to predict short-term prognosis of acute-on-chronic hepatitis B liver failure. *Chin Med J (Engl).* 2012;125(13):2272-2278.

5. Bae SK1, Yatsuhashi H, Hashimoto S, Motoyoshi Y, Ozawa E, Nagaoka S. Prediction of early HBeAg seroconversion by decreased titers of HBeAg in the serum combined with increased grades of lobular inflammation in the liver. *Med Sci Monit.* 2012;18(12):CR698.

6. Sumi A, Luo T, Zhou D, Yu B, Kong D, Kobayashi N. Time-series analysis of hepatitis A, B, C and E infections in a large Chinese city: application to prediction analysis. *Epidemiol Infect.* 2013;141(05):905-915.

7. Ren H, Li J, Yuan Z-A, Hu J-Y, Yu Y, Lu Y-H. The development of a combined mathematical model to forecast the incidence of hepatitis E in Shanghai, China. *BMC Infect Dis.* 2013;13(1):421.

8. Fabien Lavocat, Paul Dény, Christian Pichoud, Nasser Al Hawajri, Kathryn Kitrinos, Katyna Borroto-Esoda, et al. Similar evolution of hepatitis B virus quasispecies in patients with incomplete adefovir response receiving tenofovir/emtricitabine combination or tenofovir monotherapy. *J Hepatol.* 2013;59(4):684-695.

9. Lin CW, Lin CC, Mo LR, Chang CY, Perng DS, Hsu CC, et al. Heavy alcohol consumption increases the incidence of hepatocellular carcinoma in hepatitis B virus-related cirrhosis. *J Hepatol.* 2013;58(4):730-735.

10. Richardson AM, Lidbury BA. Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data. *BMC bioinformatics.* 2013;14(1):206.

11. Ye QH1, Qin LX, Forgues M, He P, Kim JW, Peng AC, et al. Predicting hepatitis B virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nat Med.* 2003;9(4):416-423.

12. Zipkin EF, Jennelle CS, Cooch EG. A primer on the application of Markov chains to the study of wildlife disease dynamics. *Methods Ecol Evol.* 2010;1(2):192-198.

13. Bharucha-Reid AT. *Elements of the Theory of Markov Processes and their Applications*. Toronto: Courier Dover Publications; 2012.

14. Lange K. *Numerical analysis for statisticians.* Tokyo: Springer; 2010.

15. Cowen ME, Chartrand M, Weitzel WF. A Markov model of the natural history of prostate cancer. *J Clin Epidemiol.* 1994;47(1):3-21.

16. Xiao Y, Tang S, Zhou Y, Smith RJ, Wu J, Wang N. Predicting the HIV/AIDS epidemic and measuring the effect of mobility in mainland China. *J Theor Biol.* 2013;317:271-285.

17. Peng Z1, Bao C, Zhao Y, Yi H, Xia L, Yu H, et al. Weighted Markov chains for forecasting and analysis in Incidence of infectious diseases in jiangsu Province, China. *J Biomed Res.* 2010;24(3):207-214.

18. Mode CJ, Sleeman CK. *Stochastic Processes in Epidemiology: HIV/AIDS, Other Infectious Diseases, and Computers.* Paris: OECD Publishing; 2000.

19. Trapman P. *On stochastic models for the spread of infections*. Amesterdam: Pieter Trapman. 2006.

20. Johnson RA, Wichern DW, Education P. *Applied multivariate statistical analysis.* Vol 4. New Jersey: Prentice hall Englewood Cliffs; 1992.

21. Prado R, West M. *Time series: modeling, computation, and inference.* Boca Raton: Chapman & Hall/CRC; 2010.

22. Brockwell PJ, Davis RA. *Introduction to time series and forecasting.* Vol 1. Boca Raton: Chapman & Hall/CRC; 2002.

23. Safee S, Ahmad S. Comparing the univariate modeling techniques and box-jenkin for measuring of climate index in sitiawan, Malaysia. *J Math Comput Sci.* 2014;4(2):350-362.

24. Chatfield C, Yar M. Holt-Winters forecasting: some practical issues. *The Statistician.* 1988:129-140.

25. Chatfield C. *The analysis of time series: an introduction.* Boca Raton: Chapman & Hall/CRC; 2013.

26. Hyndman R, Koehler AB, Ord JK, Snyder RD. *Forecasting with exponential smoothing: the state space approach.* Berlin: Springer; 2008.

27. Mitavskiy B, Cannings C. Estimating the ratios of the stationary distribution values for Markov chains modeling evolutionary algorithms. *Evol Comput.* 2009;17(3):343-377.