Epidemiol. Infect. (2013), **141**, 905–915. © Cambridge University Press 2012
doi:10.1017/S095026881200146X

Time-series analysis of hepatitis A, B, C and E infections in a large Chinese city: application to prediction analysis

A. SUMI[1]*, T. LUO[2], D. ZHOU[3], B. YU[2], D. KONG[2] AND N. KOBAYASHI[1]

[1] *Department of Hygiene, Sapporo Medical University School of Medicine, Sapporo, Hokkaido, Japan*
[2] *Department of Infectious Diseases Prevention & Control, Wuhan Centers for Disease Prevention & Control, Wuhan, Hubei, China*
[3] *Wuhan Centers for Disease Prevention & Control, Wuhan, Hubei, China*

Received 18 January 2012; Final revision 15 May 2012; Accepted 8 June 2012; first published online 20 July 2012

SUMMARY

Viral hepatitis is recognized as one of the most frequently reported diseases, and especially in China, acute and chronic liver disease due to viral hepatitis has been a major public health problem. The present study aimed to analyse and predict surveillance data of infections of hepatitis A, B, C and E in Wuhan, China, by the method of time-series analysis (MemCalc, Suwa-Trast, Japan). On the basis of spectral analysis, fundamental modes explaining the underlying variation of the data for the years 2004–2008 were assigned. The model was calculated using the fundamental modes and the underlying variation of the data reproduced well. An extension of the model to the year 2009 could predict the data quantitatively. Our study suggests that the present method will allow us to model the temporal pattern of epidemics of viral hepatitis much more effectively than using the artificial neural network, which has been used previously.

Key words: Epidemics, hepatitis, infectious disease control, preventable diseases, statistics.

INTRODUCTION

Worldwide, viral hepatitis is recognized as one of the most frequently reported diseases, and especially in China, acute and chronic liver disease due to viral hepatitis have been a major public health problem [1]. Accordingly, much effort to predict and prevent viral hepatitis infection has been expended through, for example, infectious disease surveillance, vaccinations, and various theoretical and experimental research [2]. Among these, there has been considerable interest in interpreting the mechanisms of the epidemic of viral hepatitis infection with mathematical models [3–6] and time-series analysis [7–9].

Recently, Guan *et al*. [9] used an artificial neural network to predict the incidence of hepatitis A in a large Chinese city. However, the artificial neural network is not easy to control the procedure of prediction. In addition, in China, epidemiological patterns of viral hepatitis infections vary across the country due to its diversity regarding socioeconomic conditions, ethnicity, and culture [10]. It is necessary to establish a new method of time-series analysis applicable to any time-series without restriction. In our previous study, a new analysis method which was composed of spectral analysis based on the maximum entropy method (MEM) in the frequency domain and nonlinear least squares method (LSM) in the time

* Author for correspondence: Dr A. Sumi, Department of Hygiene, Sapporo Medical University School of Medicine, S-1, W-17, Chuo-ku, Sapporo, 060-8556, Japan.
(Email: sumi@sapmed.ac.jp)

domain were proposed [11–13], and successfully used for the prediction of infectious disease epidemics [14–16]. Further, in the present study, our method of prediction analysis was applied to the surveillance data of hepatitis A, B, C and E infections in Wuhan, which is the capital city of Hubei province in central of China. Wuhan introduced mass vaccinations for hepatitis A and B from 1992 and 1986, respectively. For hepatitis C and E, no vaccine is currently available. Wuhan has accumulated good quality data on infectious diseases through its surveillance programme. Using this surveillance data on hepatitis A, B, C and E infections a model to explain the time period 2004–2008 was constructed, the model was then used to predict the time period 2009.

MATERIAL AND METHOD

Material

In the present study, prediction analysis was conducted for the surveillance data of monthly numbers of hepatitis A, B, C and E cases per 100 000 in Wuhan. The monthly data were gathered over 72 months from January 2004 to December 2009 (72 data points).

The monthly data are shown in Fig. 1*a–d* for hepatitis A, B, C and E, respectively. Each month's data was divided into an analysis range (January 2004–December 2008) and a prediction range (January–December 2009). In Figure 1, the small vertical line in the left-hand panels indicates the boundary between the analysis range (January 2004–December 2008) and the prediction range (January–December 2009).

The monthly data of hepatitis A, B, C and E infections were reported by all hospitals in Wuhan and were collected by the National Infectious Disease Reporting System, Wuhan Center for Disease Prevention and Control, China. The diagnoses of viral hepatitis infection were conducted according to the National Diagnosis Criteria. The subtypes for hepatitis A, B, C and E were divided by serological test.

Time-series analysis

Time series $x(t)$, where $t = \text{time}$, is assumed to be composed of systematic and fluctuating parts [17]:

$$x(t) = \text{systematic part} + \text{fluctuating part}. \qquad (1)$$

The systematic part in equation (1) is regarded as an underlying variation of $x(t)$, which corresponds to the predictable part [18]. The fluctuating part in equation (1), resulting from a nonlinear dynamic mechanism existing behind the data and/or undeterministic components such as noise, is obtainable as a residual time-series in which the underlying part is subtracted from the original time-series. The extrapolation curve of the underlying part can be used for prediction.

A key point is the estimation of underlying variation. The underlying variation can be determined by applying the nonlinear LSM to $x(t)$. Then, the underlying variation is assumed to be described as the function $x_{\mathrm{UV}}(t)$ given by a linear combination of sine and cosine functions,

$$x_{\mathrm{UV}}(t) = a_0 + \sum_{n=1}^{S} \{a_n \sin(2\pi f_n t) + b_n \cos(2\pi f_n t)\}, \qquad (2)$$

where f_n ($= 1/T_n$, T_n: its period) is the frequency of the nth periodic component, a_n and b_n the amplitudes of the nth component, S the total number of components, and a_0 a constant which indicates the average value of the time-series.

The optimum function of equation (2) can be determined through the nonlinear LSM for fitting analysis in the time domain. Linearization of this nonlinearity is achieved by using the frequency f_n estimated by spectral analysis based on MEM. MEM is considered to have a high degree of resolution of spectral estimation. As a result, the method of spectral analysis enabled us to make an extremely precise determination of periodic structures of time-series including a short data sequence. A formulation of MEM spectral analysis is given in the Appendix.

An outline of the analysis procedure for prediction analysis is described as follows. The details of the procedure for the method are described in our previous work [14].

(1) *Setting up time-series data for the analysis.* Equal sampling time intervals are chosen, lack of data compensated for, outliers corrected, logarithm transformation performed, and removal of long-term trend of data performed, if necessary.

(2) *Determination of f_n (MEM spectral analysis).* A spectral analysis based on MEM is conducted, and the power spectral density (PSD) is obtained. The values of f_n in equation (2) are determined by the position of the spectral peak in the PSD.

(3) *Determination of S (assignment of fundamental modes).* From the PSD, fundamental modes

https://doi.org/10.1017/S095026881200146X Published online by Cambridge University Press

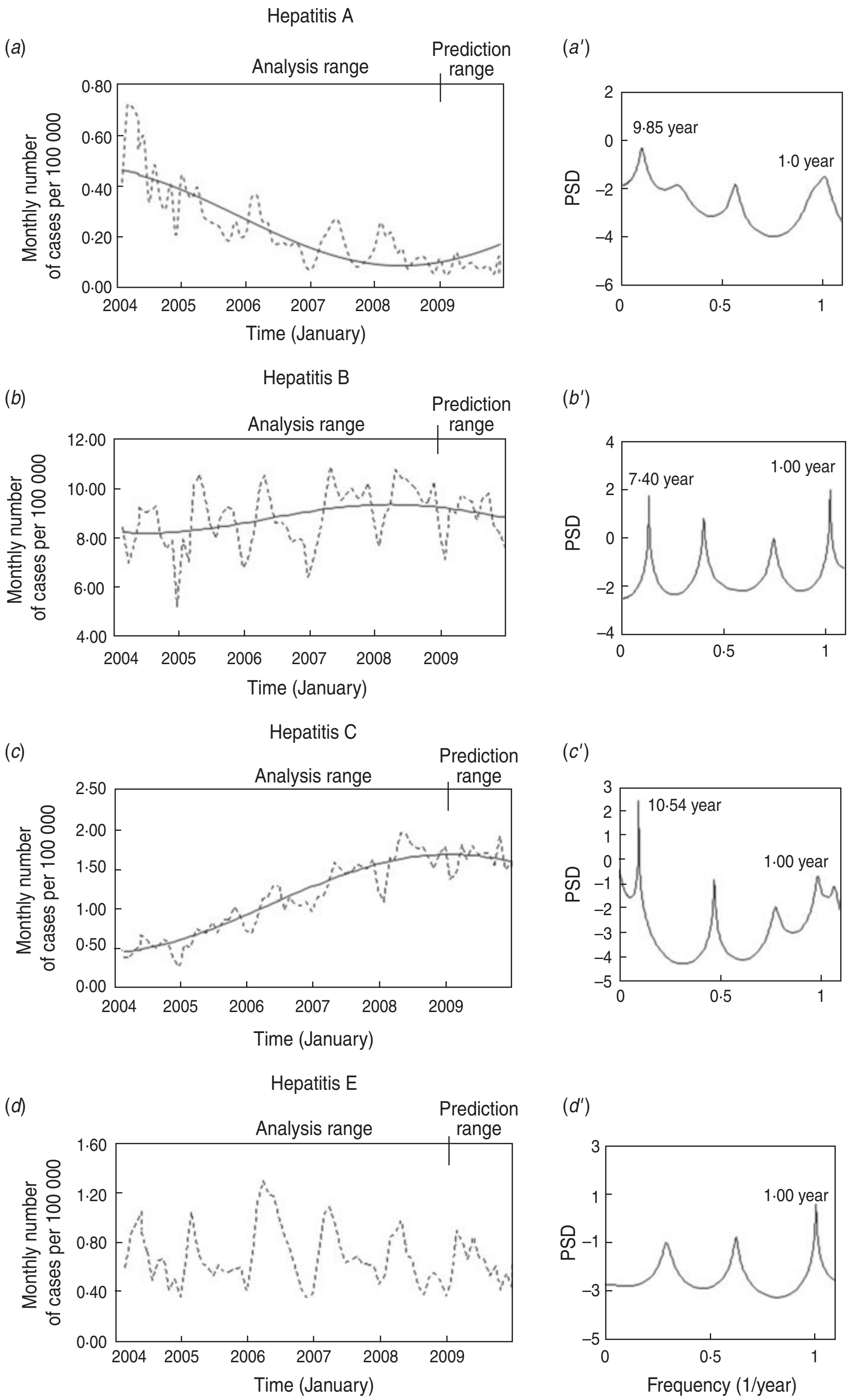

Fig. 1. Monthly data of viral hepatitis infection in Wuhan, China from 2004 to 2009, long-term trend of the data, and power spectral density (PSD) of the data. (*a–d*) The data (– – –) and the long-term trend of the data (—). (*a*) Hepatitis A, (*b*) hepatitis B, (*c*) hepatitis C, and (*d*) hepatitis E. (*a′–d′*) The PSD: (*a′*) hepatitis A, (*b′*) hepatitis B, (*c′*) hepatitis C, and (*d′*) hepatitis E.

constructing underlying variation $x_{\mathrm{UV}}(t)$ [equation (2)] of time-series data are determined. For the assignment of fundamental modes, the 'contribution ratio' is defined to indicate a criterion for the evaluation of the adequacy of $x_{\mathrm{UV}}(t)$ of time-series data [15]. The assignment of

https://doi.org/10.1017/S095026881200146X Published online by Cambridge University Press

fundamental modes results in the determination of the value of S in equation (2).

The contribution ratio against the value of number of periodic mode, S, is described as

$$\frac{\sum_{i=1}^{S} A_i^2}{Q_j}, \qquad j = \begin{cases} A\text{: analysis range} \\ P\text{: prediction range} \end{cases}$$

where A_i indicate the amplitude of the ith mode constituting the least squares fitting (LSF) curve, and Q_A and Q_P the total powers of the original time-series in the analysis and prediction ranges, respectively. An outline of the contribution ratio is described in the Appendix.

(4) *Determination of* a_0, a_n *and* b_n (*LSF analysis*). By using the estimated values of S and f_n, the optimum values of parameters a_0, a_n and b_n ($n=1, 2, \ldots, S$) in equation (2) are exactly determined with LSM. As a result, the optimum LSF curve for time-series data is obtained.

(5) *Prediction analysis*. The LSF curve is extended from the analysis range to the prediction range of time-series data, and future values are indicated quantitatively.

RESULTS

Monthly data of hepatitis A, B, C and E infections

In the monthly data for hepatitis A, B, C, and E infections (Fig. 1*a*–*d*), all data indicate a 1-year cycle. For hepatitis A (Fig. 1*a*), a large decrease in trend of the data is observed. In the case of hepatitis B (Fig. 1*b*), two peaks in spring and summer are superimposed on a 1-year cycle. The pattern of hepatitis C (Fig. 1*c*) shows a large increasing trend of the data. The pattern of hepatitis E (Fig. 1*d*) clearly indicates large peaks in spring months with small peaks apparent during summer/autumn months in the annual cycle.

Setting up the monthly data of hepatitis A, B, C and E infections for analysis

The PSDs, $P(f)$'s [f (1/year): frequency], for the data of hepatitis A, B, C and E infections in Figure 1 were calculated, and the results obtained are shown in Figure 1(*a*′–*d*′) for hepatitis A, B, C and E, respectively. Regarding hepatitis A, B and C (Fig. 1*a*′–*c*′), the longest periods appear as prominent peaks corresponding to a position longer than the length of the disease infection data in the analysis range (5 years): i.e. a 9·85-year period for hepatitis A (Fig. 1*a*′), a 7·40-year period for hepatitis B (Fig. 1*b*′) and a 10·54-year period for hepatitis C (Fig. 1*c*′). With these long-term periodic modes for hepatitis A, B and C, the long-term trend of each disease infection data was estimated by calculating the LSF curve with equation (2); the results are shown in Fig. 1(*a*–*c*). As seen in the figure's panels, LSF curves reproduce well the long-term trend in the disease infection data. The LSF curves were removed from the disease infection data, and the residual data were obtained (Fig. 2*a*–*c*). By using the residual data for hepatitis A, B and C (Fig. 2*a*–*c*) and the original data for hepatitis E (Fig. 2*d*), the prediction analysis was conducted.

Spectral analysis

PSDs for the data in the analysis range (Fig. 2) were calculated, and the semi-log scale plots ($f \leqslant 4{\cdot}5$) are shown in Figure 3. As seen in the figure's panels, several well-defined spectral lines are observed in each PSD. Ten spectral peak-frequency modes were selected, and these are summarized with the corresponding periods and intensities (powers) of the spectral peaks in Table 1.

For all PSDs (Fig. 3), prominent spectral peaks were observed at $f=1{\cdot}0$ ($=f_1$) corresponding to a 1-year period, i.e. the seasonal cycle of disease epidemics. In the case of hepatitis A (Fig. 3*a*), it is notable that dominant spectral line is also observed at the position of the 4·07-year period, which is longer than a 1-year period. For hepatitis B, C and E (Fig. 3*b*–*d*), dominant spectral lines are observed around $f=2{\cdot}0$ (6 months), which is a cause of much interest as to whether the 6-month periodic mode takes its origin from the harmonics of f_1, or the seasonal variation, or a superimposition of both.

Assignment of fundamental modes

The contribution ratio against the value of number of periodic modes, S, were calculated with the periodic modes listed in Table 1, and the results obtained are shown in Fig. 4(*a*–*d*) for hepatitis A, B, C and E, respectively.

For hepatitis A (Fig. 4*a*) the contribution ratio in the prediction range increases in the region of S from 1 to 3 as well as in the case of the analysis range. At $S=3$, the value of S in the prediction range has the largest value. Thus, three fundamental modes at $S=3$ (4·07, 1·82, 1·00 years) were assigned. The

https://doi.org/10.1017/S095026881200146X Published online by Cambridge University Press

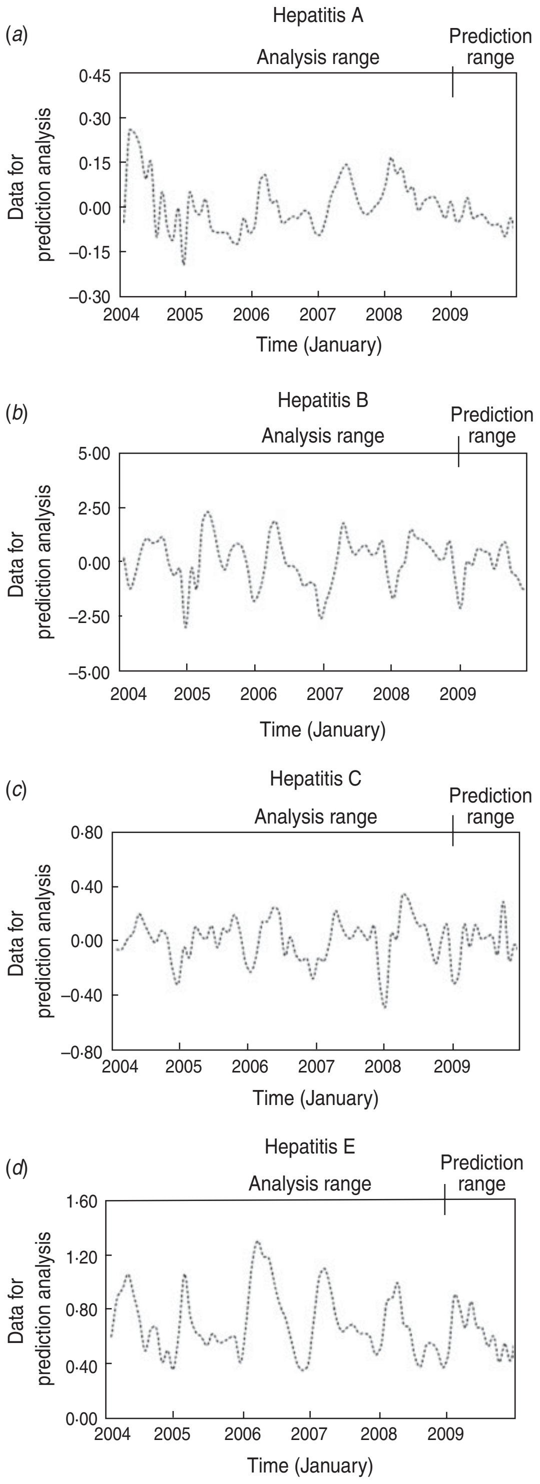

Fig. 2. The data for prediction analysis. (*a*) Hepatitis A, (*b*) hepatitis B, (*c*) hepatitis C, and (*d*) hepatitis E. Small vertical lines (|) indicate the boundary between the analysis and prediction ranges.

values of the contribution ratio at $S=3$ in the analysis and prediction ranges were 0·693 and 0·841, respectively.

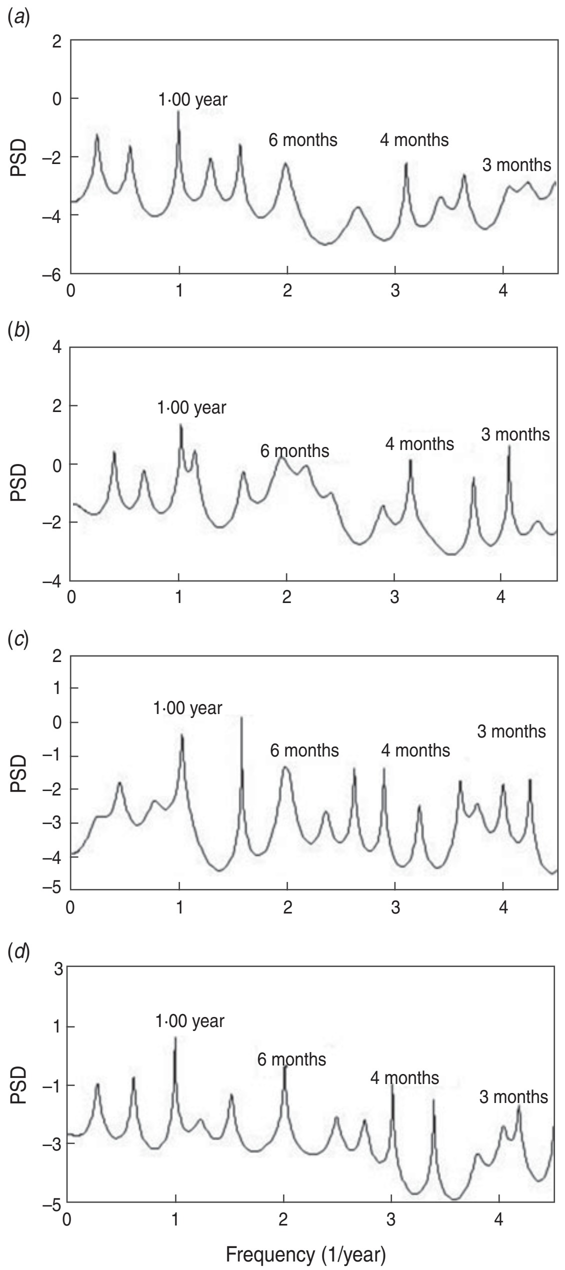

Fig. 3. Power spectral density (PSD) obtained by maximum entropy method spectral analysis ($f<4{\cdot}5$). (*a*) Hepatitis A, (*b*) hepatitis B, (*c*) hepatitis C, and (*d*) hepatitis E.

For hepatitis B (Fig. 4*b*) the contribution ratio in the prediction range increases in the region of S from 1 to 7. The contribution ratio at $S=7$ in the prediction range has the largest value, and is almost the same as that in the analysis range. Thus, seven periodic modes could be assigned as fundamental modes for the LSF curve at $S=7$ (2·64, 1·52, 1·00, 0·89, 0·52, 0·46, 0·20 years). The values of the contribution ratio at $S=7$ in the analysis and prediction ranges were 0·862 and 0·854, respectively.

https://doi.org/10.1017/S095026881200146X Published online by Cambridge University Press

Table 1. *Characteristics of the ten dominant spectral peaks shown in Figure 3*

Hepatitis A			Hepatitis B			Hepatitis C			Hepatitis E		
f	Period (yr)	Power	*f*	Period (yr)	Power	*f*	Period (yr)	Power	*f*	Period (yr)	Power
0·25	4·07[a]	0·001793	0·38	2·64[a]	0·08400	0·46	2·19[a]	0·00127	0·29	3·42[a]	0·00543
0·55	1·82[a]	0·000837	0·66	1·52[a]	0·03770	0·78	1·29[a]	0·00069	0·63	1·60	0·00532
1·00	1·01[a]	0·002472	1·00	1·00[a]	0·38640	1·03	0·97[a]	0·00916	1·01	0·99[a]	0·02591
1·29	0·78	0·000461	1·13	0·89[a]	0·12650	1·58	0·63[a]	0·0033	1·24	0·81	0·00106
1·56	0·64	0·00066	1·57	0·64	0·03570	1·98	0·51[a]	0·00359	1·53	0·66	0·00234
1·98	0·51	0·000411	1·93	0·52[a]	0·23000	2·62	0·38[a]	0·00075	2·02	0·50[a]	0·00816
4·21	0·24	0·000182	2·15	0·46[a]	0·10160	2·89	0·35[a]	0·00056	2·49	0·40	0·00076
4·49	0·22	0·000218	3·12	0·32	0·03220	3·59	0·28[a]	0·00064	2·75	0·36	0·00046
4·69	0·21	0·00026	4·03	0·25	0·03200	3·99	0·25[a]	0·00056	3·01	0·33	0·00098
5·33	0·19	0·00023	5·02	0·20[a]	0·04470	4·92	0·20[a]	0·00138	4·18	0·24	0·00067

[a] The assigned fundamental modes.

For hepatitis C (Fig. 4*c*), the contribution ratio in the prediction range increases gradually as the value of S increases, but is smaller than the contribution ratio in the analysis range for all S values. Thus, the optimum value of S which is suitable to use for calculating the LSF curve could not be found. For convenience, the ten periodic modes for the LSF curve at $S=10$ (2·19, 1·29, 1·00, 0·63, 0·51, 0·38, 0·35, 0·28, 0·25, 0·20 years) were assigned. The values of the contribution ratio at $S=10$ in the analysis and prediction ranges were 0·93 and 0·70, respectively.

For hepatitis E (Fig. 4*d*), the contribution ratio in the prediction range retains large values around 0·8, and the contribution ratio at $S=3$ in the prediction range is almost the same as the contribution ratio in the analysis range. Thus, three fundamental modes at $S=3$ (3·42, 1·00, 0·50 years) were assigned. The values of the contribution ratio at $S=3$ in the analysis and prediction ranges were 0·840 and 0·863, respectively. The values of period, amplitude and time of acrophase for the fundamental modes for each disease are listed in Table 2.

Prediction analysis

With the fundamental modes listed in Table 2, the optimum LSF curve for each data in the analysis range (January 2004–December 2008) was calculated. By extending it to the prediction range (January–December 2009), future values were indicated quantitatively. The results obtained are shown in Figure 5.

In the case of hepatitis A (Fig. 5*a*), the optimum LSF curve in the analysis range reproduces basically a 1-year cycle with large peaks in spring. Thus, fundamental modes (Table 2*a*) were confirmed to be appropriate. In the prediction range, the optimum LSF curve also reproduces well the peaks in spring.

Regarding hepatitis B and C (Fig. 5*b*, *c*), each LSF curve in the analysis range adequately reproduces a 1-year cycle and shorter-term fluctuations than the 1-year cycle of the original data. Thus, the usefulness of the fundamental periods (Table 2*b*, *c*) was confirmed. On the other hand, in the prediction range, the LSF curve for hepatitis B (Fig. 5*b*) does not reproduce the two large peaks in spring and summer of the original data. With respect to hepatitis C (Fig. 5*c*), the LSF curve in the prediction range reproduces well the spring peak of the original data, but the peak in autumn is not well reproduced.

In the case of hepatitis E (Fig. 5*d*), the LSF curve in the analysis range reproduces large peaks in spring and mild occurrences during summer/autumn months. In the prediction range, the LSF curve also reproduces well the temporal pattern of the original data.

For the LSF curve in the prediction range of each disease (Fig. 5), almost all data points of the data fit within 95% confidence intervals tested by t distribution, $x(t)=Y(t) \pm t_{0.05}s$, where $Y(t)$ is given by the estimated regression line by plotting $x_{\mathrm{UV}}(t)$ [equation (2)] against the original data in the prediction range, where s indicates standard error.

DISCUSSION

In the present study, prediction analysis for data of hepatitis A, B, C and E infections in Wuhan, China, was conducted, by investigating periodic structures of the data with MEM spectral analysis.

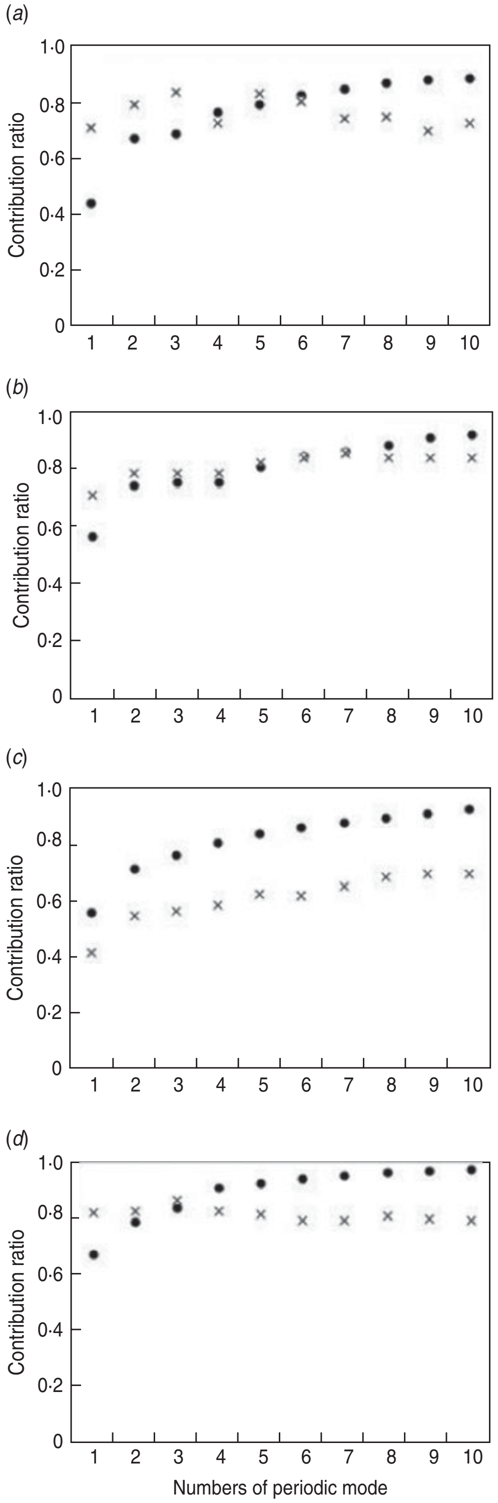

Fig. 4. Contribution ratios in the analysis (●) and prediction (×) ranges. (*a*) Hepatitis A, (*b*) hepatitis B, (*c*) hepatitis C, and (*d*) hepatitis E.

Table 2. *Parameters of fundamental modes*

Period (years)	Amplitude (A_i)	Time of acrophase
(*a*) Hepatitis A		
4·07	0·06	17 Dec. 2007
1·82	1·82	30 Apr. 2004
1·00	0·06	14 Apr. 2004
(*b*) Hepatitis B		
2·64	0·47	7 July 2005
1·52	0·28	9 July 2004
1·00	1·05	17 June 2004
0·89	0·27	14 June 2004
0·52	0·72	11 Mar. 2004
0·46	0·38	8 June 2004
0·20	0·33	12 Apr. 2004
(*c*) Hepatitis C		
2·19	0·05	18 Jan. 2006
1·29	0·03	18 Jan. 2006
1·00	0·14	25 Oct. 2004
0·63	0·06	7 July 2004
0·51	0·09	9 Apr. 2004
0·38	0·03	14 May 2004
0·35	0·04	3 Feb. 2004
0·28	0·04	18 Mar. 2004
0·25	0·03	6 Feb. 2004
0·20	0·05	23 Mar. 2004
(*d*) Hepatitis E		
3·42	0·08	3 Nov. 2006
1·00	0·22	21 Apr. 2004
0·50	0·11	25 Mar. 2004

Spectral analysis has progressed through several stages since the turn of the century. Schuster's periodogram was the first technique of modern spectral analysis. The periodogram, however, reconstructs a time-series from a sum of its Fourier components based on a prior knowledge of the fundamental period of the original time-series data. However, such knowledge is rarely obtained in practice. Therefore, the periodogram is available only for extremely restrictive cases, i.e. harmonic time-series, in which theoretically exact solutions are given. Another important approach of time-series analysis is the autoregressive (AR) model [19], which is a special case of the linear filter model, including the sophisticated version such as the autoregressive moving average (ARMA) model and the seasonal autoregressive-integrated moving average (SARIMA) model [20, 21]. However, the AR model using random noise has a weakness for interpreting the multiple periodic structures with characteristic fluctuations caused by nonlinear dynamics. On the other hand, a method of spectral analysis conducted in the present study, which is based on MEM, is applicable to any time-series without any restriction [22]. As a result,

https://doi.org/10.1017/S095026881200146X Published online by Cambridge University Press

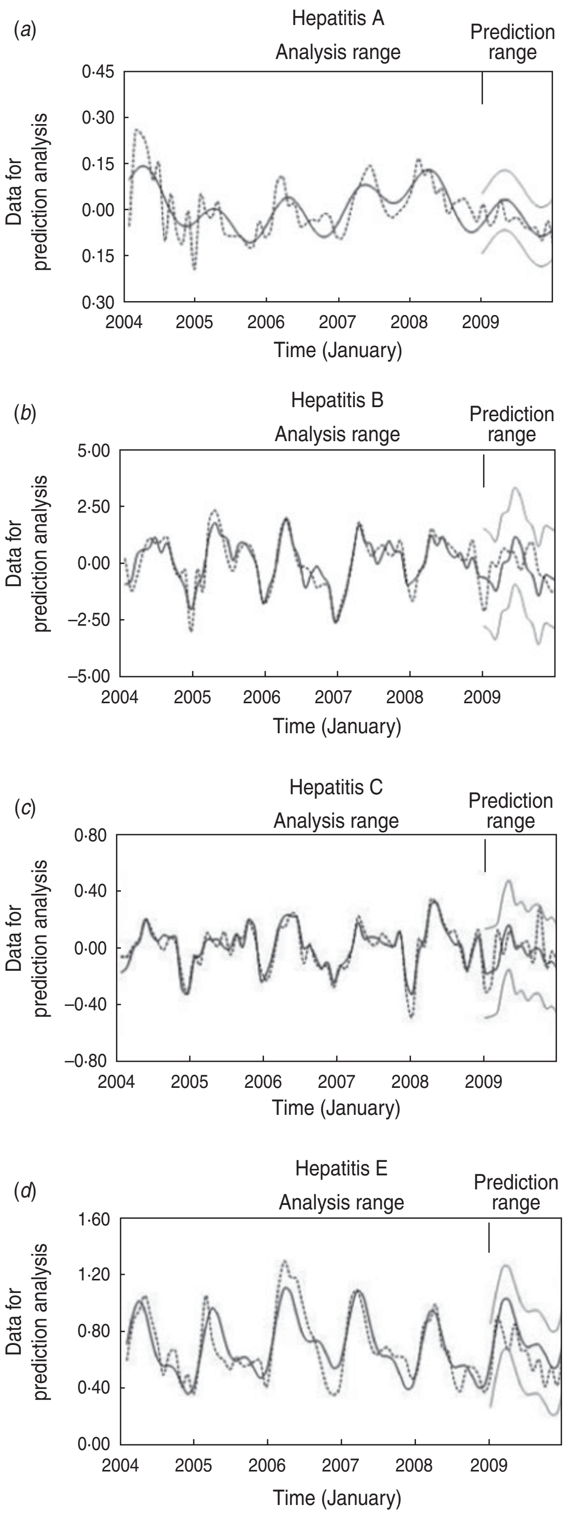

Fig. 5. Comparison of the optimum least squares fitting curve (—) with the data for prediction analysis (– – –) in the analysis range (January 2004–December 2008) and the prediction range (January–December 2009): (*a*) Hepatitis A, (*b*) hepatitis B, (*c*) hepatitis C, and (*d*) hepatitis E. Grey lines indicate 95 % confidence intervals. Small vertical lines (|) indicate the boundary between the analysis and prediction ranges.

the present method of analysis can be used to investigate periodic structures of hepatitis A, B, C and E infection data in detail, and valuable knowledge on periodic structures of the data (Fig. 3, Table 1) were obtained.

In the present study, for all diseases, the fundamental modes constructing the underlying variation of the data in the analysis range $x_{\mathrm{UV}}(t)$ [equation (2)] were successfully assigned (Table 2). As a result, for each disease in the analysis range, $x_{\mathrm{UV}}(t)$ with good fitness to the original data $x(t)$ was obtained (Fig. 5). By extending $x_{\mathrm{UV}}(t)$ from the analysis range to the prediction range, in the cases of hepatitis A and E, the original data $x(t)$ in the prediction range could be quantitatively indicated by extension of $x_{\mathrm{UV}}(t)$. This predictability for hepatitis A and E (Fig. 5*a*, *d*) was interpreted by the following explanation: the fundamental modes for the original data in the analysis range (Table 2*a*, *d*) construct the periodic structure of the underlying variations of the original data in both analysis and prediction ranges. The unpredictability observed for hepatitis B and C (Fig. 5*b*, *c*) might be because of the temporal periodic structures of the disease, and the fundamental modes in the analysis range (Table 2*b*, *c*) are not preserved in the prediction range. This may mean that the original data for hepatitis B and C include a large amount of fluctuations corresponding to the 'fluctuating part' in equation (1). For the origin of the fluctuation, in the case of hepatitis B (Fig. 5*b*), two reasons are considered; first, the fluctuations resulting from nonlinear dynamics [5], and second, the fluctuations resulting from random noise because of a large number of chronic cases [23]. The chronic cases of hepatitis B result from the fact that disease infections are mainly transmitted by perinatal exposure in Wuhan. In addition, horizontal infection due to sexual and iatrogenic transmission may also play an important role in disease infections [24].

With respect to hepatitis C, chronic cases might result in fluctuations of the original data as well as for the case of hepatitis B [25]. Chronic cases of hepatitis C are related to the fact that, in Wuhan, transmission of the disease has been caused by blood transfusion, sharing syringe needles infected by drug abusers, and other sources of iatrogenic infection. A systematic review of the prevalence of hepatitis C infection among injecting drug users, reported that the epidemic was most severe in the southern inland province, especially in Hubei province (where Wuhan is the capital city) [24]. With respect to the large

https://doi.org/10.1017/S095026881200146X Published online by Cambridge University Press

increasing trend of hepatitis C infections (Fig. 1*c*), corresponding to a 10·54-year period (Fig. 1*c*′), this may result from the fact that the test for hepatitis C virus antibody has been conducted in high-risk groups of the disease such as injecting drug users.

For hepatitis A, so far, researchers have suggested that, in Europe and the USA, prior to and immediately following World War II, hepatitis rates were high, and nationwide epidemics of hepatitis A occurred at 6- to 10-year intervals [26, 27]. It is considered that this temporal pattern trend has been changed by socioeconomic factors, i.e. improved sanitation and hygienic standards [23]. With respect to the route of transmission of hepatitis A in Wuhan, disease incidence occurs regularly, because of environmental pollution and poor hygiene [28]. Thus, the fundamental mode of 4·07 years assigned for hepatitis A in the present study (Table 2) may be explained by socioeconomic factors that promote hepatitis A virus transmission in Wuhan. For the results of long-term trends of hepatitis A (Fig. 1*a*), the large decreasing trend of disease infections corresponding to the 9·85-year period (Fig. 1*a*′) may be a result of the introduction of vaccination against hepatitis A in 1992 in Wuhan.

In the case of hepatitis E, a 6-month periodic mode assigned as the fundamental mode (Table 2) interprets the seasonal variations of the disease, with a peak in spring and a smaller peak in summer (Fig. 2*d*). In Wuhan, transmission of hepatitis E is clearly related to the faecal–oral route, usually through contaminated drinking water, and zoonotic transmission from pigs [29]. The decades-long surveillance conducted in China also suggested that pigs constitute a major reservoir and source of hepatitis E infections [24]. For the spring peak of hepatitis E, it was reported that the disease virus transmission in Southwest England was closely related to the presence of pig [30, 31]. On the other hand, in Wuhan, the infection of hepatitis E virus was higher in urban areas, where people do not live in close proximity to pigs. Thus, with respect to the spring peak in Wuhan (Fig. 2*d*), hepatitis E virus infections may be spread through faecal–oral transmission with contamination of drinking water during the Chinese New Year holiday at the beginning of February, with the spring peak occurring after the incubation period of 2–10 weeks. As well as the case of the spring peak, the seasonal variation in the summer/autumn peak can be considered to be due to the faecal–oral route, usually through contamination of water supplies [32, 33]. Based on the studies reported so far, the summer/autumn peak usually occurs in those parts of the world, where heavy rains occur or monsoon conditions are present; high rates of disease have persisted through rainy seasons followed by a significant decrease in the number of cases and the ending of the epidemic. In Wuhan, the rainy season starts from the end of May and ends in the early July. Thus, it is possible that mild occurrences of hepatitis E during summer/autumn relate to the rainy season.

The present method of time-series analysis, which is applicable to any time-series without any restriction, can be successfully used for prediction analysis even where the data for hepatitis B and C infections include a large amount of fluctuations due to chronic cases. In conclusion, it is anticipated that the present method of time-series analysis consisting of MEM spectral analysis and LSM will contribute to further development in the field of prediction analysis of epidemics of viral hepatitis.

APPENDIX

MEM spectral analysis

The PSD obtained by MEM spectral analysis for time-series data under analysis, with an equal sampling interval Δt ($=1$ month, in the present study), can be calculated from

$$P_m(f)=\frac{P_m\Delta t}{\left|1+\sum_{k=-m}^{m}\gamma_{m,k}\exp\left[-i2\pi fk\Delta t\right]\right|^2}, \qquad \text{(A1)}$$

where P_m is the output power of a prediction-error filter of order m and $\gamma_{m,k}$ the corresponding filter coefficient, $m=0,\ 1,\ 2,\ \ldots,\ M$; where M is the optimum filter order. P_m and $\gamma_{m,k}$ are determined by solving the following Yule–Walker equations with the use of Burg's procedure.

Determination of the value of the 'contribution ratio'

The determination of S in equation (2) is made via the following procedure. Based on the result of periods estimated by MEM spectral analysis, we must assign fundamental modes f_n that construct a periodic structure of $x_{\mathrm{UV}}(t)$ [equation (2)]. Then, we investigate the contribution of ten dominant periods estimated by MEM spectral analysis of the LSF curve in the analysis and prediction ranges: (*a*) the LSF curve in the analysis range is calculated with the variation S, by the ten modes being added to the LSF curve one by

https://doi.org/10.1017/S095026881200146X Published online by Cambridge University Press

one in the order of magnitude of the power of the spectral peak frequency, (*b*) the LSF curve calculated with each S is extended to the underlying variations in the prediction range, and (*c*) the evaluation of the LSF curve is performed. The procedure (*c*) is divided into four steps [(*c*1), (*c*2), (*c*3), (*c*4)]. In procedure (*c*1), the power of each periodic mode is evaluated by the square of amplitude, A_i^2, of the ith mode constituting the LSF curve. In procedures (*c*2) and (*c*3), we estimate R_j corresponding to the power of time-series which is obtained by subtracting the LSF curve from the original time-series. As a result, the total powers of the original time-series in the analysis and prediction ranges (Q_A and Q_P, respectively) are obtained by

$$Q_j = \sum_{i=1}^{S} A_i^2 + R_j \qquad j = \begin{cases} A\text{: analysis range} \\ P\text{: prediction range} \end{cases} \qquad \text{(A2)}$$

When both sides of equation (A2) are divided by Q_j, we obtain the following normalized relation:

$$\frac{\sum_{i=1}^{S} A_i^2}{Q_j} + \frac{R_j}{Q_j} = 1 \qquad \begin{cases} A\text{: analysis range} \\ P\text{: prediction range} \end{cases}$$

where $\sum_{i=1}^{S} A_i^2/Q_j$ and R_j/Q_j correspond to the contribution of $\sum_{i=1}^{S} A_i^2$ and R_j to Q_j, respectively. Then, in procedure (*c*4), we define the first term of the left-hand side of equation (A3) the 'contribution ratio', which means the contribution $\sum_{i=1}^{S} A_i^2$ normalized by Q_j. If $\sum_{i=1}^{S} A_i^2/Q_j$ in the first term becomes large, then the second term R_j/Q_j becomes small.

ACKNOWLEDGEMENTS

This work was supported by a Grant-in-Aid for Scientific Research (Grant no. 22406017) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Lee L, Lv Jun.** Public health in China: history and contemporary challenges. In: Beaglehole R, Bonita R, eds. *Global Public Health: A New Era*, 2nd edn. Oxford: Oxford University Press, 2009. pp. 185–207.
2. **Wen YM, Xu ZY, Melnick JL (eds).** *Viral Hepatitis in China: Problems and Control Strategies* (Monographs in Virology, vol. 19). Basel: Karger, 1992.
3. **Zhao S, Xu Z, Lu Y.** A mathematical model of hepatitis B virus transmission and its application for vaccination strategy in China. *International Journal of Epidemiology* 2000; **29**: 744–752.
4. **Dickinson JA, Wun YT, Wong SL.** Modelling death rates for carriers of hepatitis B. *Epidemiology and Infection* 2002; **128**: 83–92.
5. **Zou L, Zhang W, Ruan S.** Modelling the transmission dynamics and control of hepatitis B virus in China. *Journal of Theoretical Biology* 2010; **262**: 330–338.
6. **Goldstein ST, *et al.*** A mathematical model to estimate global hepatitis B disease burden and vaccination impact. *International Journal of Epidemiology* 2005; **34**: 1329–1339.
7. **Rolfhamre P, Ekdahl K.** An evaluation and comparison of three commonly used statistical models for automatic detection of outbreaks in epidemiological data of communicable diseases. *Epidemiology and Infection* 2006; **134**: 863–871.
8. **Naumova EN, *et al.*** Seasonality in six enterically transmitted diseases and ambient temperature. *Epidemiology and Infection* 2007; **135**: 281–292.
9. **Guan P, Huang DS, Zhou BS.** Forecasting model for the incidence of hepatitis A based on artificial neural network. *World Journal of Gastroenterology* 2004; **10**: 3579–3582.
10. **Tanaka M, *et al.*** Hepatitis B and C virus infection and hepatocellular carcinoma in China: a review of epidemiology and control measures. *Journal of Epidemiology* 2011; **21**: 401–416.
11. **Luo T, *et al.*** Study on the effect of measles control programmes on periodic structures of disease epidemics in a large Chinese city. *Epidemiology and Infection* 2011; **139**: 257–264.
12. **Ohtomo K, *et al.*** Relationship of cholera incidence to El Ninõ and solar activity elucidated by time-series analysis. *Epidemiology and Infection* 2010; **138**: 99–107.
13. **Sumi A, *et al.*** Study of the effect of vaccination of periodic structures of measles epidemics in Japan. *Microbiology and Immunology* 2007; **51**: 805–814.
14. **Sumi A, Kamo K.** MEM spectral analysis for predicting influenza epidemics in Japan. *Environmental Health and Preventive Medicine* 2012; **17**: 98–108.
15. **Sumi A, *et al.*** Prediction analysis for measles epidemics. *Japanese Journal of Applied Physics* 2003; **42**: 7611–7620.
16. **Sumi A, *et al.*** Predicting the incidence of human campylobacteriosis in Finland with time series analysis. *Acta Pathologica, Microbiologica et Immunologica Scandinavica* 2009; **117**: 614–622.
17. **Armitage P, Berry G, Matthews JNS.** *Statistical Method in Medical Reseach*, 4th edn. Oxford: Blackwell Science, 2002.
18. **Populis A.** *Probability, Random Variables, and Stochastic Processes*, 3rd edn. New York: MacGraw-Hill, 1991.
19. **Benschop J, *et al.*** Temporal and longitudinal analysis of Danish Swine Salmonellosis Control Programme data: implications for surveillance. *Epidemiology and Infection* 2008; **136**: 1511–1520.
20. **José MV, Bishop RF.** Scaling properties and systematic patterns in epidemiology of rotavirus infection. *Philosophical Transactions of the Royal Society of*

https://doi.org/10.1017/S095026881200146X Published online by Cambridge University Press

London, Series B: Biological Sciences 2003; **358**: 1625–1641.

21. **Nobre FF, *et al.*** Dynamical linear model and SARIMA: a comparison of their forecasting performance in epidemiology. *Statistics in Medicine* 2001; **20**: 3051–3069.
22. **Ohtomo N, Tanaka Y.** New method of time series analysis and 'MemCalc'. In: Saito K, Koyama A, Yoneyama K, Sawada Y, Ohtomo N, eds. *A Recent Advance in Time Series Analysis by Maximum Entropy method*. Sapporo: Hokkaido University Press, 1994, pp. 11–30.
23. **Margolis HS, Alter MJ, Hadler SC.** Viral hepatitis. In: Evans AS, Kaslow RA, eds. *Viral Infections of Humans: Epidemiology and Control*, 4th edn. New York: Plenum Medical Book Co., 1997, pp. 363–418.
24. **Lu J, *et al.*** General epidemiological parameters of viral hepatitis A, B, C, and E in six regions of China: a cross-sectional study in 2007. *PLoS ONE* 2009; **4**: e8467.
25. **National Institute of Infectious Diseases.** Hepatitis B and C in Japan as of May 2002. *Infectious Agents Surveillance Report* 2002; **23**: 163–164.
26. **Shaw FE, *et al.*** A community-wide epidemic of hepatitis in Ohio. *American Journal of Epidemiology* 1986; **123**: 1057–1065.
27. **Shapiro CN, *et al.*** Epidemiology of hepatitis A in the United States. In: Hollinger FB, Lemon SM, Margolis H, eds. *Viral Hepatitis and Liver Disease – Proceeding of the 1990 International Symposium on Viral Hepatitis and Liver Disease: Contemporary Issues and Future Prospects*. Baltimore: Williams & Wilkins, 1991, pp. 71–76.
28. **Bell BP.** Global epidemiology of hepatitis A: implications for control strategies. *10th International Symposium on Viral Hepatitis and Liver Disease*, 2002. International Medical Press (http://ec.digaden.edu.mx/moodle/moodledata/16/01medint/01enfinf/01ei/07-12/07g.pdf).
29. **Kong D, *et al.*** Analysis on epidemiological features of hepatitis A and hepatitis E in Wuhan City from 2004 to 2009 [in Chinese]. *Chinese Journal of Disease Control & Prevention* 2011; **15**: 701–704.
30. **Dalton HR, *et al.*** Autochthonous hepatitis E in southeast England: natural history, complications and seasonal variation, and hepatitis E virus IgG seroprevalence in blood donors, the elderly and patients with chronic liver disease. *European Journal of Gastroenterology & Hepatology* 2008; **20**: 784–790.
31. **Ijaz S, *et al.*** Non-travel-associated hepatitis E in England and Wales: demographic, clinical and molecular epidemiological characteristics. *Journal of Infectious Diseases* 2005; **192**: 1166–1172.
32. **Aggarwal R, Naik S.** Epidemiology of hepatitis E: current status. *Journal of Gastroenterology and Hepatology* 2009; **24**: 1484–1493.
33. **Zhuang H, *et al.*** Epidemiology of hepatitis E in China. *Gastroenterologia Japonica* 1991; **26** (Suppl. 3): 135–138.

https://doi.org/10.1017/S095026881200146X Published online by Cambridge University Press