BRIEF ARTICLES

An autoregressive integrated moving average model for short-term prediction of hepatitis C virus seropositivity among male volunteer blood donors in Karachi, Pakistan

Saeed Akhtar, Shafquat Rozi

Saeed Akhtar, Department of Community Medicine and Behavioral Sciences, Faculty of Medicine, Kuwait University, PO Box 24923, Safat 13110, Kuwait

Shafquat Rozi, Department of Community Health Sciences, Medical College, Aga Khan University, Stadium Road, Karachi 74800, Pakistan

Correspondence to: Saeed Akhtar, PhD, Department of Community Medicine and Behavioral Sciences, Faculty of Medicine, Kuwait University, PO Box 24923, Safat 13110, Kuwait. saeed.akhtar@hsc.edu.kw
Telephone: +965-2498-6542 Fax: +965-2533-8948

Abstract

AIM: To identify the stochastic autoregressive integrated moving average (ARIMA) model for short term forecasting of hepatitis C virus (HCV) seropositivity among volunteer blood donors in Karachi, Pakistan.

METHODS: Ninety-six months (1998-2005) data on HCV seropositive cases ($1000^{-1} \times month^{-1}$) among male volunteer blood donors tested at four major blood banks in Karachi, Pakistan were subjected to ARIMA modeling. Subsequently, a fitted ARIMA model was used to forecast HCV seropositive donors for 91-96 mo to contrast with observed series of the same months. To assess the forecast accuracy, the mean absolute error rate (%) between the observed and predicted HCV seroprevalence was calculated. Finally, a fitted ARIMA model was used for short-term forecasts beyond the observed series.

RESULTS: The goodness-of-fit test of the optimum ARIMA (2,1,7) model showed non-significant autocorrelations in the residuals of the model. The forecasts by ARIMA for 91-96 mo closely followed the pattern of observed series for the same months, with mean monthly absolute forecast errors (%) over 6 mo of 6.5%. The short-term forecasts beyond the observed

series adequately captured the pattern in the data and showed increasing tendency of HCV seropositivity with a mean ± SD HCV seroprevalence ($1000^{-1} \times month^{-1}$) of 24.3 ± 1.4 over the forecast interval.

CONCLUSION: To curtail HCV spread, public health authorities need to educate communities and health care providers about HCV transmission routes based on known HCV epidemiology in Pakistan and its neighboring countries. Future research may focus on factors associated with hyperendemic levels of HCV infection.

Akhtar S, Rozi S. An autoregressive integrated moving average model for short-term prediction of hepatitis C virus seropositivity among male volunteer blood donors in Karachi, Pakistan. *World J Gastroenterol* 2009; 15(13): 1607-1612 Available from: URL: http://www.wjgnet.com/1007-9327/15/1607.asp DOI: http://dx.doi.org/10.3748/wjg.15.1607

INTRODUCTION

Hepatitis C virus (HCV) infection poses a major public health problem in developing countries, including Pakistan. However, the results of prevalence studies have shown variable estimates in select groups including 1.8% to 3.0% in volunteer blood donors[1,2] and 16% to 20.5% in familial contacts of infected patients[3,4]. A community-based study in Hafizabad, Punjab, found a 6.5% HCV seroprevalence[5]. Using these estimates, Pakistan has been grouped into intermediate category with respect to burden of HCV infection[6]. Several routes have been implicated for nosocomial and community acquired

HCV infection including unsafe injections, recycling of used syringes, inadequate sterilization of surgical and dental equipment, and facial shaving by barbers[2,7].

Public health authorities in Pakistan intermittently run educational campaigns in electronic and print media to create awareness in the general population to halt HCV spread. However, in the absence of adequate HCV surveillance, the true impact of the HCV control efforts remains uncertain. Volunteer blood donors are generally considered to be the healthier segment of any community and the proportions of HCV seropositivity among them may be considered to mirror the situation in the general population[8]. We have previously reported a significant increase in HCV seroprevalence among volunteer blood donors over the past several years using data from two blood banks[2]. However, there is need to expand this HCV surveillance network countrywide to obtain more reliable and representative estimates.

Recently, mathematical models have been used to project the future HCV prevalence among intravenous drug users[9], and its impact on the future development of HCV related morbidity and mortality[10]. Modeling and forecasting HCV seropositivity among volunteer blood donors in Pakistan, and perhaps in other neighboring countries, might provide useful information for allocating resources, and re-shaping and planning future control activities[11]. This study aimed to develop a univariate time series model for HCV seropositivity ($1000^{-1} \times month^{-1}$) among volunteer blood donors attending four large blood banks. Specifically, the objective of this study was to identify the stochastic autoregressive integrated moving average (ARIMA) model for short term forecasting of HCV seropositivity ($1000^{-1} \times month^{-1}$) among volunteer blood donors in Karachi, Pakistan.

MATERIALS AND METHODS

Setting
This study was conducted in Karachi-the largest cosmopolitan city and the hub of economic activity of Pakistan. It has an estimated population of 9.3 million, accounting for approximately 10% of the total population of the country. Forty three percent of the city's population is under the age of 15 years. The population of Karachi comprises several ethnic groups defined by mother tongue, including predominantly Urdu, Sindhi, Punjabi, Pushto, and Balochi. The healthcare facilitates for the population include several small and tertiary care hospitals, both in the private and public sector.

Data
Eight-year (1998-2005) data on monthly aggregates of number of donors attending four large blood banks (blood bank I -IV) in Karachi were available for this study. These blood banks receive blood donations only from non-remunerated volunteer blood donors. Blood bank I is part of a tertiary care hospital in the private sector and receives blood donations as replacements

from friends and relatives of inpatients requiring blood transfusions. Blood banks II-IV belong to non-governmental organizations and cater for the needs of those in Karachi who need blood transfusions, including the patients with leukemia, hemophilia, thalassemia and other blood related diseases. Blood banks II-IV also receive blood donations from volunteers on an exchange basis. Prior to blood donation, each blood donor is subjected to screening for known risk factors for transfusion transmissible infections. All the blood banks follow similar criteria to receive blood donations and exclude potential donors who admit known risk factors of transfusion transmissible infections or any medical or non-medical condition associated with high risk (e.g. use of narcotic drugs, history of jaundice in the past 5 years and recent hospitalization). All four blood banks in the study use commercially available enzyme-linked immunosorbant assay kits and results are interpreted according to the manufacturer's instructions.

As noted earlier, blood donations between January 1998 and December 2005 by men aged 18-64 years were included in this evaluation. HCV serological results of consecutive blood donations from these blood banks were available from variable starting dates depending on the completed records, to assess the proportions of HCV seropositive donors.

Analytic approach
We used methods developed by Box and Jenkins to build an ARIMA time series model[12]. This model-building process is designed to take advantage of associations in the sequentially lagged relationships that usually exist in data collected periodically. The general form of the ARIMA model was

$$\Delta_1 z_t = \Phi_1 z_{t-1} + \ldots + \Phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}$$

where:

$\Delta_1 z_t$ = differenced series i.e. $z_t - z_{t-1}$

z_t = set of possible observations on the time-sequenced random variable

a_t = random shock term at time t

$\Phi_1 \ldots \Phi_p$ = autoregressive parameters of order p

$\theta_1 \ldots \theta_p$ = moving average parameters of order q

The series was subjected to Box-Cox transformation[13]. The transformed series was then differenced at the non-seasonal level and mean corrected to induce stationarity. Sample autocorrelation and partial autocorrelation functions were used to identify the ARIMA model of the appropriate order. Estimates of the model's parameters were obtained by the maximum likelihood method. Diagnostic checking included residual analysis and the Akaike Information Criterion was used to compare goodness-of-fit among ARIMA models. The final model was a result of several iterations of the identification, estimation, and checking process, and met the conventional criteria for the adequacy of the model[14].

Assessment of forecast accuracy
The last 6 observations in the data set were used for validation of the forecast accuracy of the ARIMA
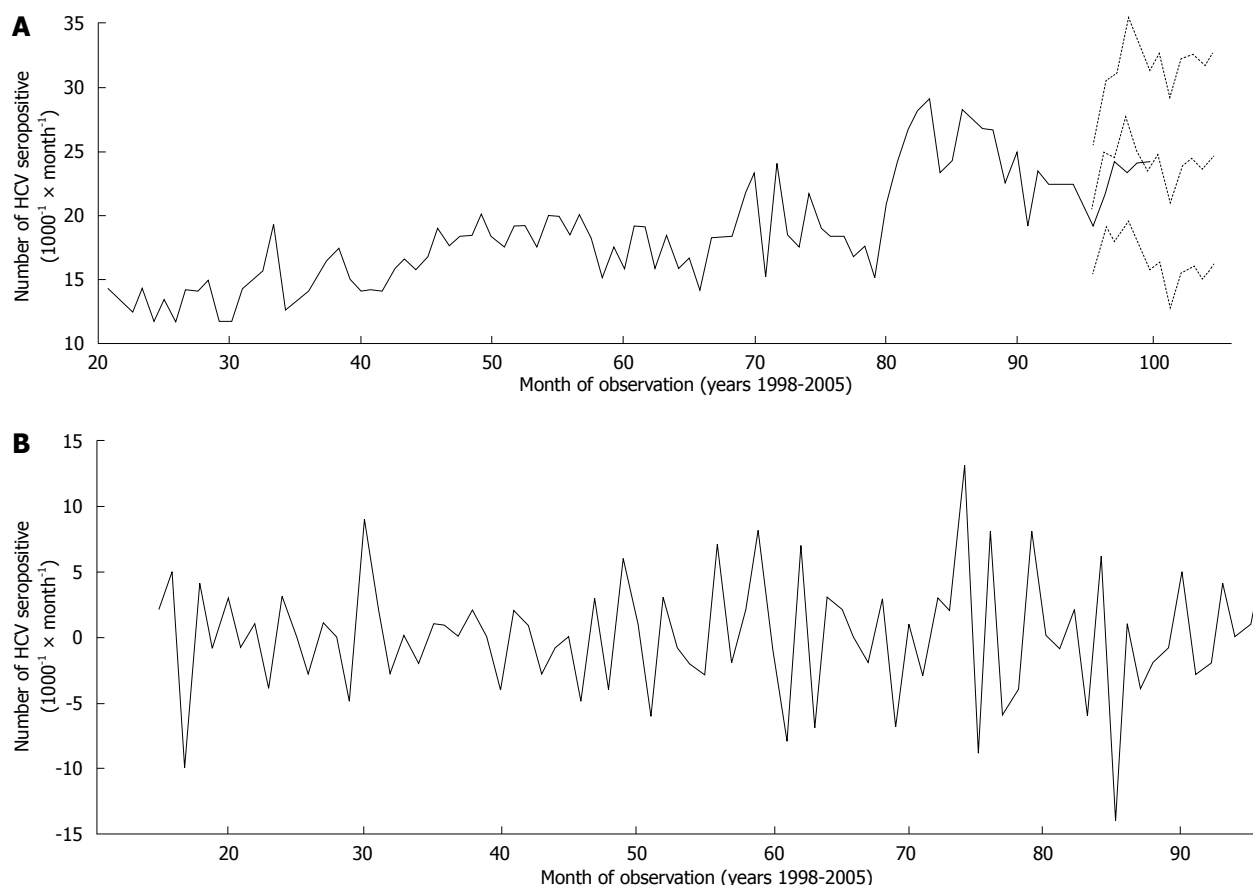
Figure 1 **Hepatitis C virus seroprevalence (1000^{-1} × month^{-1}) among volunteer male blood donors in Karachi, Pakistan 1998-2005.** A: Observed data along with forecasts; B: Transformed series.

Yr	Mean	SD	95% CI for mean		Minimum	Maximum
			Lower limit	Upper limit		
1998	13.8	1.5	12.8	14.7	12	16
1999	16.3	2.2	14.8	17.7	13	21
2000	18.5	2.2	17.1	19.9	15	22
2001	19.8	2.0	18.6	21.1	16	22
2002	19.9	3.0	18.0	21.8	15	26
2003	20.3	3.2	18.2	22.3	16	27
2004	29.3	2.6	27.6	30.9	25	33
2005	24.8	2.1	23.4	26.1	21	27
Total	20.3	5.1	19.3	21.3	12	33

Table 1 Hepatitis C virus seroprevalence (1000^{-1} × year^{-1}) among male volunteer blood donors at four large blood banks in Karachi (1998- 2005)

$F = 47.9$; $df = 7$, 88; $P < 0.001$.

model. The fitted ARIMA model was used to forecast the HCV seroprevalence (1000^{-1} × month^{-1}) for 91-96 mo (June 2005 to December 2005) to contrast with the observed series of the same months. The average forecast error at prediction interval of *m* months ($\overline{\varepsilon}_m$) was calculated as:

$$(\overline{\varepsilon}_m) = [\sum_{m=1}^{6}(y_{t=m} - \hat{y}_{t=m})/6]^{1/2}$$

Where $y_{t=m}$ and $\hat{y}_{t=m}$ denote the observed and forecast values for month $t + m$. Finally, the fitted ARIMA model was used for short term (January 2006 to June 2006) forecasts along with their 95% confidence limits beyond the observed series.

RESULTS

Descriptive analysis
The crude HCV seropositivity (1000^{-1}) among the male volunteer blood donors during the study period was 20.3 (12792/630134). The mean prevalence (1000^{-1} × month^{-1}) was 18.3 [95% confidence interval (CI): 16.8-19.9]. There was no statistically significant difference in HCV seropositivity (1000^{-1}) across various months of the years ($F = 0.201$; $P = 0.997$) (data not shown). However, a substantial variation in HCV seroprevalence (1000^{-1}) was observed across different calendar years ($F = 47.895$; $P < 0.001$) (Table 1). The observed and transformed series are presented in Figure 1.

ARIMA model
The parameters' estimates for the optimum ARIMA (2,1,7) model for the series of monthly HCV seropositive donors (1000^{-1}) are shown in Table 2. The autocorrelation and partial autocorrelation functions of the residuals showed good-fit (Figure 2). The residual plots showed small variations around the zero mean. None of these residuals had its magnitude larger than twice the standard deviation. Residuals' autocorrelations were not significantly different from zero as a set and had constant variance, thus confirming the adequacy of the model (Ljung-Box statistic = 20.4; $P = 0.433$).

The forecasts by the ARIMA (2,1,7) model for 91-96 mo (June 2005 to December 2005) using the
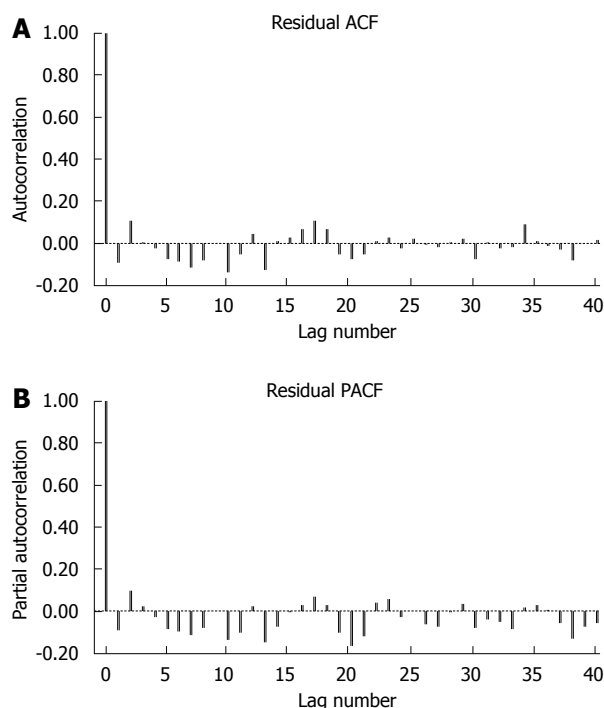
Figure 2 Residual plots for the final ARIMA (2,1,7) model of HCV seroprevalence ($1000^{-1} \times$ month^{-1}) among male volunteer blood donors in Karachi, Pakistan 1998-2005. A: Autocorrelation function; B: Partial autocorrelation function.

Table 2 Autoregressive integrated moving average model (2,1,7) of hepatitis C virus seroprevalence ($1000^{-1} \times$ month^{-1}) among male volunteer donors in Karachi, Pakistan, (January 1998-December 2005)

Parameters	Estimate	Standard error	t-ratio
Autoregressive parameter (Φ)			
Φ_2	0.67	0.15	4.5
Moving average parameter (θ)			
θ_2	-0.59	0.18	3.3
θ_3	0.49	0.15	3.3
θ_4	-0.80	0.11	7.3
θ_6	0.74	0.21	3.5
θ_7	-0.37	0.17	2.2

White noise variance = 9.74; Ljung-Box Q statistic = 20.4 (P = 0.433).

observed series of months 1-90, closely followed the pattern of observed series for the same months (Figure 1), with mean ± SD and maximum monthly absolute forecast errors (%) over 6 mo interval being 6.5% ± 3.4% and 10%, respectively. Furthermore, the short term (January 2006 to June 2006) forecasts beyond the observed series adequately captured the pattern in the data (Figure 1) and showed evidence of increasing tendency of HCV seroprevalence ($1000^{-1} \times$ month^{-1}) with the mean ± SD as 24.3 ± 1.4 over the forecast interval.

DISCUSSION

Epidemiological surveillance of communicable diseases is one of the more traditional public health activities. Time series analysis of surveillance data on prevalence and/or incidence of various infections may be helpful in developing hypotheses to explain and anticipate the dynamics of the observed phenomena and subsequently in the establishment of a quality control system and re-allocation of resources[15,16]. This method is an ecologic approach and takes advantage of the strong association in the sequentially lagged relationship that usually exists in the data collected periodically[17].

During the study period, the overall HCV seroprevalence (1000^{-1}) in volunteer blood donors was 20.3, which falls in the range of 14.9 to 38.9 known for first time blood donors in other developing countries. However, HCV seroprevalence (1000^{-1}) in this study was much higher than the 2.1 reported for developed countries[18]. The low HCV seroprevalence in resource-rich countries is attributed to safe blood transfusion, whereas, in poor regions of the world, several million people acquire HCV infection each year as a result of contaminated transfusions and the re-use of infected medical devices[18,19]. Therefore, public health practices adopted by the developed countries need to be strictly enforced in less developed countries to break the chain of transmission of HCV and other blood-borne pathogens.

Monitoring of HCV seropositivity among volunteer blood donors may provide clues about the effectiveness of control efforts of public health authorities and future trend of the proportion of HCV infected donors in Pakistan. In this paper, we used the ARIMA model on a time series of HCV seropositivity (1000^{-1}) collected monthly over a period of 96 mo on asymptomatic male volunteer blood donors from four major blood banks in Karachi. The forecasts made in a prospective manner over six months demonstrated increasing tendency of HCV seropositivity among the blood donors in this cosmopolitan city. Such a predicted increase in HCV seropositivity might result from inconsistent and naïve HCV control efforts on the part of public health officials in Pakistan. Therapeutic injections in a health-care setting have consistently been shown as a strong risk factor for HCV infection in Pakistan[2,7,20,21], and if concerted efforts by the public health authorities are not made, might continue to contribute to the increasing load of HCV infection in this and similar settings in the region. An increasing trend among first time US blood donors of 50 to 59 years of age from 1995 to 2002 has been demonstrated[22]. According to the authors, teenage children and young adults in 1960, and 1970s might have experimented with drug injection and were infected with HCV. These people entered into the 50 to 59 years age group during 1995 to 2002. However, in other age groups of donors in the same study and two other studies from US[23,24], and from other developed countries (France[25,26] and Spain[27]) have shown a decreasing trend of residual risk of HCV infection in blood donors. According to these investigators different factors could have played a role in this reduction, for instance, increased awareness about the factors associated with increased risk of HCV infection, voluntary deferral by potential high risk

donors, improvement in donor recruitment, and /or an overall decrease in HCV infection level in the general population. Such factors need to be evaluated in our population in future studies.

Results from our previous study[2], and those predicted by ARIMA model for 6 mo beyond the observed data exhibited a slightly increasing tendency of HCV seropositivity among male volunteer blood donors over the forecast period. This increasing pattern of HCV seroprevalence among these asymptomatic male volunteer donors merits further investigation of factors contributing to HCV seroprevalence in this population, which is thought to be a mirror image of the situation in the general population.

Some limitations of this study need to be taken into account when interpreting the results. Our HCV seroprevalence estimates are based on ELISA, which has sensitivity of more than 95%. However, these results do not reflect possible HCV infections that do not produce detectable seropositivity during the window period of HCV infection. The exact proportion of these HCV infected, but HCV seronegative, is not known, however, it has been argued that this figure must be very small given the use of current sero-assays[28]. Our HCV seroprevalence estimates in male volunteer donor population were based on data from a limited number of blood banks; we do not know whether they reflect the national average. The blood banks that participated in this study however, account for a substantial proportion of donations made annually in Karachi. These centers are located in large metropolitan areas where the prevalence and/or incidence of HCV may be higher than the national figures. Therefore, we think we are justified in making generalizations from our data. In conclusion, in the absence of comprehensive HCV surveillance in the general population in Pakistan and perhaps in other neighboring countries, further monitoring of HCV seropositivity in blood donors and the investigation of factors associated with hyperendemic HCV infection using multivariate ARMIA models might further expand our understanding about HCV epidemiology in this region. Furthermore, effective screening of all blood donors for HCV infection at all blood banks should be seriously considered, because one single HCV infected regular blood donor could transmit the infection to several recipients.

ACKNOWLEDGMENTS

COMMENTS

Background

Less developed countries such as Pakistan generally lack effective surveillance systems for communicable diseases. In this study, the authors used the Box-Jenkins approach to fit an autoregressive integrated moving average (ARIMA) model that might be used to monitor and predict trends in hepatitis C virus (HCV) seroprevalence in the general population using volunteer blood donors as a sentinel group. This information may be helpful to facilitate early public health responses to minimize HCV related morbidity and mortality.

Research frontiers

Developed countries have been able to control the HCV transmission in the general population by public health measures. However, such initiatives are practiced at sub-optimal level in resource-constrained countries. This problem is further compounded by the absence of effective surveillance of communicable diseases including blood-borne pathogens. Therefore, alternative methods to monitor and predict the burden of such infections are needed for rational allocation of resources.

Innovations and breakthroughs

This is the first application of an ARIMA model to monitor and predict the HCV seroprevalence in volunteer blood donors at multiple blood banks. Such data on infections with HCV and other blood-borne pathogens may mirror the situation in a setting that lacks an effective surveillance system for these infections.

Application

The fitted ARIMA model could be used for sentinel surveillance of blood-borne infections in volunteer blood donors. Therefore, the estimates for current and predicted future burden of these infections could be used by public health authorities for making rational policy decisions for control and prevention of HCV and other blood-borne pathogens in resource-constrained countries including Pakistan.

Peer review

The authors describe an effective model to predict HCV seropositivity in Pakistan.

REFERENCES

1   **Kakepoto GN**, Bhally HS, Khaliq G, Kayani N, Burney IA, Siddiqui T, Khurshid M. Epidemiology of blood-borne viruses: a study of healthy blood donors in Southern Pakistan. *Southeast Asian J Trop Med Public Health* 1996; **27**: 703-706

2   **Akhtar S**, Younus M, Adil S, Jafri SH, Hassan F. Hepatitis C virus infection in asymptomatic male volunteer blood donors in Karachi, Pakistan. *J Viral Hepat* 2004; **11**: 527-535

3   **Pasha O**, Luby SP, Khan AJ, Shah SA, McCormick JB, Fisher-Hoch SP. Household members of hepatitis C virus-infected people in Hafizabad, Pakistan: infection by injections from health care providers. *Epidemiol Infect* 1999; **123**: 515-518

4   **Akhtar S**, Moatter T, Azam SI, Rahbar MH, Adil S. Prevalence and risk factors for intrafamilial transmission of hepatitis C virus in Karachi, Pakistan. *J Viral Hepat* 2002; **9**: 309-314

5   **Luby SP**, Qamruddin K, Shah AA, Omair A, Pahsa O, Khan AJ, McCormick JB, Hoodbhouy F, Fisher-Hoch S. The relationship between therapeutic injections and high prevalence of hepatitis C infection in Hafizabad, Pakistan. *Epidemiol Infect* 1997; **119**: 349-356

6   Hepatitis C: global prevalence. *Wkly Epidemiol Rec* 1997; **72**: 341-344

7   **Bari A**, Akhtar S, Rahbar MH, Luby SP. Risk factors for hepatitis C virus infection in male adults in Rawalpindi-Islamabad, Pakistan. *Trop Med Int Health* 2001; **6**: 732-738

8   **Pillonel J**, Saura C, Couroucé AM. [Prevalence of HIV, HTLV, and hepatitis B and C viruses in blood donors in France, 1992-1996] *Transfus Clin Biol* 1998; **5**: 305-312

9   **Murray JM**, Law MG, Gao Z, Kaldor JM. The impact of behavioural changes on the prevalence of human immunodeficiency virus and hepatitis C among injecting drug users. *Int J Epidemiol* 2003; **32**: 708-714

10 **Law MG**, Dore GJ, Bath N, Thompson S, Crofts N, Dolan K, Giles W, Gow P, Kaldor J, Loveday S, Powell E, Spencer J, Wodak A. Modelling hepatitis C virus incidence, prevalence and long-term sequelae in Australia, 2001. *Int J Epidemiol* 2003; **32**: 717-724

11 **Brachman PS**. Infectious diseases--past, present, and future.

Int J Epidemiol 2003; **32**: 684-686

12   **Box GEP**, Jenkins GM. Time series analysis: forecasting and control. San Francisco: Holden Day, 1976: 181-218

13   **Box GEP**, Cox DR. An analysis of transformation (with discussion). *J R Stat Soc* 1964; **B26**: 211-252

14   **Ljung GM**, Box GEP. On a measure of lack of fit in time series models. *Biometrika* 1978; **65**: 297-303

15   **Catalano R**, Serxner S. Time series designs of potential interest to epidemiologists. *Am J Epidemiol* 1987; **126**: 724-731

16   **Kuhn L**, Davidson LL, Durkin MS. Use of Poisson regression and time series analysis for detecting changes over time in rates of child injury following a prevention program. *Am J Epidemiol* 1994; **140**: 943-955

17   **Morgenstern H**. Uses of ecologic analysis in epidemiologic research. *Am J Public Health* 1982; **72**: 1336-1344

18   **Prati D**. Transmission of hepatitis C virus by blood transfusions and other medical procedures: a global review. *J Hepatol* 2006; **45**: 607-616

19   **Alter HJ**. HCV natural history: the retrospective and prospective in perspective. *J Hepatol* 2005; **43**: 550-552

20   **Khan AJ**, Luby SP, Fikree F, Karim A, Obaid S, Dellawala S, Mirza S, Malik T, Fisher-Hoch S, McCormick JB. Unsafe injections and the transmission of hepatitis B and C in a periurban community in Pakistan. *Bull World Health Organ* 2000; **78**: 956-963

21   **Janjua NZ**, Akhtar S, Hutin YJ. Injection use in two districts of Pakistan: implications for disease prevention. *Int J Qual Health Care* 2005; **17**: 401-408

22   **Zou S**, Notari EP 4th, Stramer SL, Wahab F, Musavi F, Dodd RY. Patterns of age- and sex-specific prevalence of major blood-borne infections in United States blood donors, 1995 to 2002: American Red Cross blood donor study. *Transfusion* 2004; **44**: 1640-1647

23   **Dodd RY**, Notari EP 4th, Stramer SL. Current prevalence and incidence of infectious disease markers and estimated window-period risk in the American Red Cross blood donor population. *Transfusion* 2002; **42**: 975-979

24   **Glynn SA**, Kleinman SH, Schreiber GB, Busch MP, Wright DJ, Smith JW, Nass CC, Williams AE. Trends in incidence and prevalence of major transfusion-transmissible viral infections in US blood donors, 1991 to 1996. *Retrovirus Epidemiology Donor Study (REDS) JAMA* 2000; **284**: 229-235

25   **Pillonel J**, Laperche S, Saura C, Desenclos JC, Couroucé AM. Trends in residual risk of transfusion-transmitted viral infections in France between 1992 and 2000. *Transfusion* 2002; **42**: 980-988

26   **Velati C**, Romanò L, Baruffi L, Pappalettera M, Carreri V, Zanetti AR. Residual risk of transfusion-transmitted HCV and HIV infections by antibody-screened blood in Italy. *Transfusion* 2002; **42**: 989-993

27   **Alvarez M**, Oyonarte S, Rodríguez PM, Hernández JM. Estimated risk of transfusion-transmitted viral infections in Spain. *Transfusion* 2002; **42**: 994-998

28   **Kleinman S**, Alter H, Busch M, Holland P, Tegtmeier G, Nelles M, Lee S, Page E, Wilber J, Polito A. Increased detection of hepatitis C virus (HCV)-infected blood donors by a multiple-antigen HCV enzyme immunoassay. *Transfusion* 1992; **32**: 805-813