scientific reports

OPEN

Pollutants-mediated viral hepatitis in different types: assessment of different algorithms and time series models

Shengfei Pei[1], Li Yang[2], Huixia Gao[2], Yuzhen Liu[2], Erhei Dai[2], Fumin Feng[1]✉ & Jianhua Lu[2]✉

The escalating frequency of environmental pollution incidents has raised significant concerns regarding the potential health impacts of pollutant fluctuations. Consequently, a comprehensive study on the role of pollutants in the prevalence of viral hepatitis is indispensable for the advancement of innovative prevention strategies. Monthly incidence rates of viral hepatitis from 2005 to 2020 were sourced from the Chinese Center for Disease Control and Prevention Infectious Disease Surveillance Information System. Pollution data spanning 2014–2020 were obtained from the National Oceanic and Atmospheric Administration (NOAA), encompassing pollutants such as CO, NO2, and O3. Time series analysis models, including seasonal auto-regressive integrated moving average (SARIMA), Holt-Winters model, and Generalized Additive Model (GAM), were employed to explore prediction and synergistic effects related to viral hepatitis. Spearman correlation analysis was utilized to identify pollutants suitable for inclusion in these models. Concurrently, machine learning (ML) algorithms were leveraged to refine the prediction of environmental pollutant levels. Finally, a weighted quantile sum (WQS) regression framework was developed to evaluate the singular and combined impacts of pollutants on viral hepatitis cases across different demographics, age groups, and environmental strata. The incidence of viral hepatitis in Beijing exhibited a declining trend, primarily characterized by HBV and HCV types. In predicting hepatitis prevalence trends, the Holt-Winters additive seasonal model outperformed the SARIMA multiplicative model ((1,1,0) (2,1,0) $_{[12]}$). In the prediction of environmental pollutants, the SVM model demonstrated superior performance over the GPR model, particularly with Polynomial and Besseldot kernel functions. The combined pollutant risk effect on viral hepatitis was quantified as βWQS (95% CI) = 0.066 (0.018, 0.114). Among different groups, $PM_{2.5}$ emerged as the most sensitive risk factor, notably impacting patients with HCV and HEV, as well as individuals aged 35–64. CO predominantly affected HAV patients, showing a risk effect of βWQS (95% CI) = − 0.0355 (− 0.0695, − 0.0016). Lower levels of $PM_{2.5}$ and PM_{10} were associated with heightened risk of viral hepatitis incidence with a lag of five months, whereas elevated levels of $PM_{2.5}$ (100–120 μg/m^3) and CO correlated with increased hepatitis incidence risk with a lag of six months. The Holt-Winters model outperformed the SARIMA model in predicting the incidence of viral hepatitis. Among machine learning algorithms, SVM and GPR models demonstrated superior performance for analyzing pollutant data. Patients infected with HAV and HEV were primarily influenced by PM_{10} and CO, whereas SO_2 and $PM_{2.5}$ significantly impacted others. Individuals aged 35–64 years appeared particularly susceptible to these pollutants. Mixed pollutant exposures were found to affect the development of viral hepatitis with a notable lag of 5–6 months. These findings underscore the importance of long-term monitoring of pollutants in relation to viral hepatitis incidence.

Keywords Viral hepatitis, Time series, Pollutants, Machine learning, Weighted quantile sum

[1]School of Public Health, North China University of Science of Technology, Tangshan 062310, Hebei, China. [2]Hebei Key Laboratory of Immune Mechanism of Major Infectious Diseases and New Technology of Diagnosis and Treatment, The Fifth Hospital of Shijiazhuang, Shijiazhuang, China. ✉email: fm_feng@sina.com; 13323219965@163.com

Viral hepatitis refers to liver inflammation caused by infection with one of five known viruses: hepatitis A, B, C, D, and E [1,2]. This condition poses a significant global public health challenge, affecting billions worldwide and contributing to high rates of morbidity and mortality. Hepatitis A and E typically follow a self-limiting course with full recovery, whereas hepatitis B and C often progress to chronic infection and are associated with severe health outcomes. Historical records trace the prevalence of hepatitis back to ancient times, with documented outbreaks dating back 5000 years ago in China and descriptions of jaundice recorded by Hippocrates on the island of Sássos in the fifth century BC[3]. Viral hepatitis causes over 1.4 million deaths annually[4]. In a multicenter international study across 161 countries, the prevalence of hepatitis B virus (HBV) surface antigen (HBsAg) was reported at 3.61%[5]. Despite declines in the disease burden of HBV and HCV infections globally over the past three decades, HBV remains prevalent in China[6]. Consequently, viral hepatitis has emerged as a top global health priority, prompting the implementation of extensive public health policies.

To effectively inform health policies aimed at preventing viral hepatitis, accurate prediction of its trends is paramount. Research in Iran has identified the Holt Exponential Smoothing (HES) model as highly accurate in forecasting HBV incidence[7]. However, comprehensive predictive studies for viral hepatitis remain limited. Existing literature predominantly focuses on clinical and virological factors, often overlooking environmental influences. For instance, a study in Spain demonstrated that each additional rainy day increased the risk of contracting hepatitis A two weeks later (IRR = 1.03, 95% CI = 1.01–1.05)[8]. Additionally, Chen et al.[9] found a correlation between $PM_{2.5}$ exposure and hepatitis progression to hepatocellular carcinoma, though research on the synergistic effects of pollutants with hepatitis infection remains scarce.

This study aims to investigate the epidemiological characteristics of viral hepatitis of viral hepatitis, develop predictive models using various methods, and explore the singular, multiple, and interactive effects of pollutants. Specifically, our objectives are to: (a) construct and evaluate prediction models using diverse methodologies; (b) explore the single and multiple effects of pollutants across different groups; (c) analyze pollutant interactions over lagging timeframes.

Patients and methods

Overview of the study area

Beijing, situated in northern China, covers a land area of 16,410.54 square kilometers. It is centrally located at approximately 116°20′ east longitude and 39°56′ north latitude. Beijing experiences a warm temperate semi-humid and semi-arid monsoon climate, characterized by hot and rainy summers and cold and dry winters. Administratively, the city comprises 16 districts and serves as the capital of the People's Republic of China.

Data source

Data on all reported cases of viral hepatitis in Beijing from 2005 to 2020 were sourced from the public health science data center website (https://www.phsciencedata.cn/). This dataset includes information on the incidence and morbidity of various types of viral hepatitis such as HAV, HBV, HCV, HDV, HEV, and unclassified hepatitis. Diagnosis of all patients followed the criteria outlined in the viral hepatitis management guidelines issued by the Ministry of Health of the People's Republic of China. Ethical approval for this study was obtained from the China Center for Disease Control and Prevention. To ensure confidentiality, viral hepatitis data were analyzed anonymously. Given that viral hepatitis is classified as a statutory infectious disease under national mandatory surveillance, informed consent was not required. Monthly pollutions information (2014–2020) were sourced from the National Oceanic and Atmospheric Administration (NOAA) (https://www.noaa.gov/) encompassing parameters such as AQI, $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2 and O_3.

Time series analysis of single and multiple interaction

This study employed three models for time series analysis. The SARIMA and Holt-Winters models were primarily used for predicting the incidence trends of viral hepatitis. The Holt-Winters exponential smoothing model is effective in smoothing out random fluctuations and assigns varying weights to data across cycles, thereby enhancing the accuracy of future trend predictions[10]. Holt-Winters' additive model has the following expression:

$$\begin{aligned}
\hat{y}_{t+h/t} &= l_t + hb_t + s_{t-m+h},\\
l_t &= \alpha\left(y_t - s_{t-m}\right) + (1-\alpha)(l_{t-1} + b_{t-1}),\\
b_t &= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1},\\
s_t &= \gamma\left(y_t - l_{t-1} - b_{t-1}\right) + (1-\gamma)s_{t-m}.
\end{aligned}$$

where, $0 \le \alpha \le 1$, $0 \le \beta \le 1$, $0 \le \gamma \le 1 - \alpha$. s_{t-m+h} is the seasonal term. α, β, and γ are the smoothing parameters. m is seasonal periods, and h is the predicted step size.

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model decomposes the observed values into three parts: residuals, seasonal features, and true trends[11]. The SARIMA (p, d, q) (P, D, Q) s model can be expressed as follows:

$$\begin{gathered}\Phi_p(L)A_P(L^s)\Delta^d\Delta_s^D y_t = \Theta_q(L)B_Q(L^s)\varepsilon_t,\\ \Phi_p(L) = 1 - \varphi_1 L - \varphi_2 L - \cdots - \varphi_p L^p,\\ A_P(L^s) = 1 - \alpha_1 L^s - \alpha_2 L^{2s} - \cdots - \alpha_P L^{Ps},\\ \Theta_q(L) = 1 + \theta_1 L + \theta_2 L + \cdots + \theta_q L^q,\\ B_Q(L^s) = 1 + \beta_1 L^s + \beta_2 L^{2s} + \cdots + \beta_Q L^{Qs},\\ \Delta_s y_t = (1 - L^s)y_t = y_t - y_{t-s},\\ \Delta_s = 1 - L^s,\\ \varepsilon_t : WN\left(0, \sigma^2\right)\end{gathered}$$

where, Δ and Δ_s denote non-seasonal and seasonal differences, respectively. φ, Φ, θ and Θ are the parameters of the model, ε_t is white noise with independent and identical distribution[12].

Following this, Spearman correlation analysis was used to identify relevant pollutants. Subsequently, the GAM generalized additive model (GAMs) was used to explore the interaction of pollutant factors on the prevalence of viral hepatitis[13]. The following model formula are as followed:

$$\log[E(Y_t)] = \alpha_1 + s(X_1, X_2) + \Sigma s(X_t)$$

α_1 is the intercept; X_1 and X_2 indicate two interaction pollutants; *s ()* indicates penalized spline function. *s* (X_1, X_2) is a spline function of the interaction between the parameters X_1 and X_2 (X_1 and X_2 are all 5–6 months lagged variables.). $\Sigma s(X_t)$ are the factors of non-interaction pollutants.

Machine learning training process

To predict viral hepatitis across different age groups and subtypes, various machine learning (ML) algorithms were employed, and the results compared. The modeling utilized data from 2014 to 2018 for training set and data from 2019 to 2020 for testing, with both sets undergoing ten-fold cross-validation. The Gaussian Process Regression (GPR) model operates by defining a Gaussian process to model the distribution of functions, followed by Bayesian inference in function space[14]. Four kernel function algorithms—Rbf, Polynomial, Laplace, and Bessel—were employed in the GPR model for comparison. The support vector regression (SVR) algorithms were also utilized, which map input features to a higher dimensional space, maximizing the margin between classes[15]. The SVR model compared four kernel function algorithms: Linear, Polynomial, Radial and Sigmoid. This study used R4.3.1 package e1071 and kernlab to construct SVR and GPR models, respectively. We use pollutants as predictor variables in the model of the ML algorithm. Subsequently, we consider the overall incidence of the population, the incidence among different age groups, and the incidence among different types of viral hepatitis as outcome variables. This allows us to investigate the sensitivity of different populations to air pollutants in terms of disease incidence.

Single pollution and weighted quantile sum (WQS) statistical analyses

The WQS regression model serves to evaluate the combined effects of multiple exposure variables on a specified outcome. Each exposure variable is assigned a weight within the model to quantify its influence on the outcome variable[16]. Initially, this study employs the WQS model to identify pollutants significantly impacting the incidence rate of viral hepatitis across various age groups and subtypes. To assess the cumulative impact of simultaneous exposure to multiple pollutants and discern individual contributions of each pollutant, a "mixtures" approach via WQS regression analysis was utilized. Concurrently, epidemiological data was stratified into different air quality categories based on Beijing's AQI, distinguishing between pollution and good air quality levels. Within varying environmental quality states, the WQS regression model was applied to analyze how different pollutants influence the incidence and mortality of viral hepatitis.

Results

Demographic characteristics

From Table 1, the incidence of viral hepatitis in Beijing between 2005 and 2020 exhibited a general declining trend, with a notable short-term surge observed from 2016 to 2018. Conversely, the mortality rate displayed an increasing trend, peaking at 0.77 per 100,000 in 2011. Predominantly, HBV and HCV subtypes accounted for approximately 86.25% of cases, while HDV cases were rare, totaling only three. The seasonal distribution indicated spring and summer epidemics. Among age groups, individuals aged 35–64 years constituted the majority at 51.23%, followed by those aged 15–34 years at 31.38%.

The analysis of time series model results

Comparing the predicted graphs from Fig. 1A, B, it can be observed that the Holt-Winters model outperforms the SARIMA model in time periods. In Table S1, the Deviation indicator reveals that the Holt-Winters model demonstrates a relatively minor discrepancy compared to the SARIMA model in predicting outcomes for the year 2019. However, the Holt-Winters model exhibits a notable advantage in its predictions for 2020. In Table S2, the parameters for the Holt-Winters additive model are determined as α = 0.44, β = 0.09, γ = 1, while the SARIMA multiplicative model is specified as SARIMA (1,1,0) (2,1,0) $_{[12]}$. Despite comparing metrics such as RMSE, it was found that there is little discernible difference in the performance of the two models.

Model prediction comparisons

Figure S1 showed illustrates the results of Spearman's correlation analysis, revealing positive associations between five pollutants—$PM_{2.5}$, PM_{10}, SO_2, CO and NO_2—and the prevalence of viral hepatitis. Notably, $PM_{2.5}$

Characteristic		0–14	15–34	35–64	≥ 65	Total	Incidence (10^{-5}%)	Mortality (10^{-5}%)
		No of hepatitis cases (%)						
Year	2005	128 (1.31%)	3793 (38.85%)	4369 (44.75%)	1473 (15.09%)	9763	45.33	0.18
	2006	154 (1.20%)	4925 (38.38%)	5817 (45.34%)	1935 (15.08%)	12,831	59.58	0.59
	2007	103 (1.09%)	3492 (36.93%)	4482 (47.40%)	1378 (14.57%)	9455	43.90	0.38
	2008	61 (0.86%)	2404 (34.00%)	3423 (48.41%)	1183 (16.73%)	7071	32.83	0.32
	2009	43 (0.71%)	1800 (29.74%)	3097 (51.17%)	1112 (18.37%)	6052	28.10	0.67
	2010	46 (0.86%)	1374 (25.56%)	2884 (53.65%)	1072 (19.94%)	5376	24.96	0.69
	2011	30 (0.59%)	1295 (25.61%)	2847 (56.31%)	884 (17.48%)	5056	23.48	0.77
	2012	17 (0.41%)	1097 (26.41%)	2352 (56.63%)	687 (16.54%)	4153	19.28	0.50
	2013	16 (0.47%)	902 (26.25%)	1942 (56.52%)	576 (16.76%)	3436	15.95	0.72
	2014	8 (0.26%)	763 (24.94%)	1779 (58.16%)	509 (16.64%)	3059	14.20	0.40
	2015	14 (0.47%)	737 (24.78%)	1712 (57.57%)	511 (17.18%)	2974	13.81	0.42
	2016	9 (0.31%)	727 (25.20%)	1635 (56.67%)	514 (17.82%)	2885	13.40	0.48
	2017	9 (0.28%)	948 (29.07%)	1771 (54.31%)	533 (16.34%)	3261	15.14	0.48
	2018	7 (0.20%)	952 (26.70%)	1983 (55.61%)	624 (17.50%)	3566	16.56	0.38
	2019	9 (0.30%)	668 (22.27%)	1728 (57.62%)	594 (19.81%)	2999	13.93	0.39
	2020	5 (0.24%)	479 (23.23%)	1209 (58.63%)	369 (17.90%)	2062	9.57	0.39
Classifications	HAV	113 (4.25%)	839 (31.55%)	1288 (48.44%)	419 (15.76%)	2659		
	HBV	343 (0.63%)	20,487 (37.40%)	27,194 (49.64%)	6754 (12.33%)	54,778		
	HCV	112 (0.63%)	3203 (18.13%)	9216 (52.15%)	5140 (29.09%)	17,671		
	HDV	0 (0.00%)	0 (0.00%)	2 (66.67%)	1 (33.33%)	3		
	HEV	19 (0.32%)	851 (14.24%)	3766 (63.03%)	1339 (22.41%)	5975		
	Unclassified hepatitis	72 (2.47%)	976 (33.50%)	1564 (53.69%)	301 (10.33%)	2913		
Seasons	Spring (Mar–May)	176 (0.75%)	7486 (31.84%)	11,928 (50.74%)	3919 (16.67%)	23,509		
	Summer (Jun–Aug)	219 (1.06%)	6668 (32.38%)	10,420 (50.61%)	3283 (15.94%)	20,590		
	Autumn (Sep–Nov)	134 (0.70%)	6074 (31.65%)	9749 (50.79%)	3236 (16.86%)	19,193		
	Winter (Dec–Feb)	130 (0.63%)	6128 (29.59%)	10,933 (52.80%)	3516 (16.98%)	20,707		
Total		659 (0.78%)	26,356 (31.38%)	43,030 (51.23%)	13,954 (16.61%)	83,999		

Table 1. Distribution of viral hepatitis cases by age, types and season groups in Beijing, China, 2005–2020.

shows a significant cross-correlation with both PM_{10} and CO ($r = 0.84$, $P < 0.001$). Table 2 compares four kernel algorithms of GPR, indicating relatively better predictive performance for HCV across different genotypes (R^2test ∈ [0.087, 0.202]). Similarly, among age groups, individuals aged 35 and above exhibit more accurate predictions (R^2test ∈ [0.024, 0.150]). The Besseldot kernel function within the GPR model demonstrates superior predictive capability. Table 3 evaluates four kernel algorithms of SVM, highlighting HBV as having better predictive outcomes across genotypes (R^2test ∈ [0.215, 0.303]). Additionally, individuals aged 35 and above show enhanced prediction accuracy (R^2test ∈ [0.010, 0.132]). The Polynomial kernel function proves advantageous within the SVM framework. Overall, SVM demonstrates superior predictive performance compared to GPR across the evaluated metrics, underscoring its efficacy in modeling the relationships between pollutants, genotypes, age groups, and viral hepatitis development.

Assess the combined association between multiple pollutions exposures and viral hepatitis

Table S3 presents the comprehensive sensitivity analysis, indicating that the combined effect of the five pollutants on viral hepatitis is βWQS (95% CI) = 0.066 (0.018, 0.114). Among different subtypes, pollutants demonstrate significant adverse effects on HAV, HCV, and HEV. Across different age groups, except for the 0–14 age group, pollutants show notable adverse effects. Subsequently, based on the results of the overall sensitivity analyses, the relevant key factors were initially screened. From Table 4, focusing on individual pollutant effects, $PM_{2.5}$ emerges as the primary risk factor for viral hepatitis overall, with a risk effect of βWQS (95% CI) = − 0.0050 (− 0.0089, − 0.0013). Among different subgroups, $PM_{2.5}$ stands out as the most sensitive risk factor, particularly impacting HCV and HEV patients and individuals aged 35–64. SO_2 primarily affects HCV patients and individuals aged 35–64, with risk effects of βWQS (95% CI) = 0.0022 (0.0004, 0.0040) and βWQS (95% CI) = 0.0043 (0.0005, 0.0080), respectively. CO mainly impacts HAV patients, with a risk effect of βWQS (95% CI) = − 0.0355 (− 0.0695, − 0.0016). NO_2 primarily affects individuals aged 0–14, while PM_{10} influences HEV patients. In terms of combined pollutant effects, pollutants mainly affect HCV patients and individuals aged 35–64 (with risk effects of βWQS (95% CI) = 0.0342 (0.0210, 0.0474) and βWQS (95% CI) = 0.0453 (0.0153, 0.1556), respectively).

Regarding environmental pollution periods, as illustrated by Fig. S2, SO_2 and CO are key pollutants influencing the onset and mortality of viral hepatitis. During polluted periods (Fig. S2C), SO_2 and $PM_{2.5}$ predominantly affect onset, whereas during periods of good environmental conditions (Fig. S2A), SO2 and PM2.5 are primary factors. Similarly, for mortality during polluted periods (Fig. S2D), CO and SO_2 play critical roles, while during good environmental periods (Fig. S2B), CO and $PM_{2.5}$ are significant influencers.

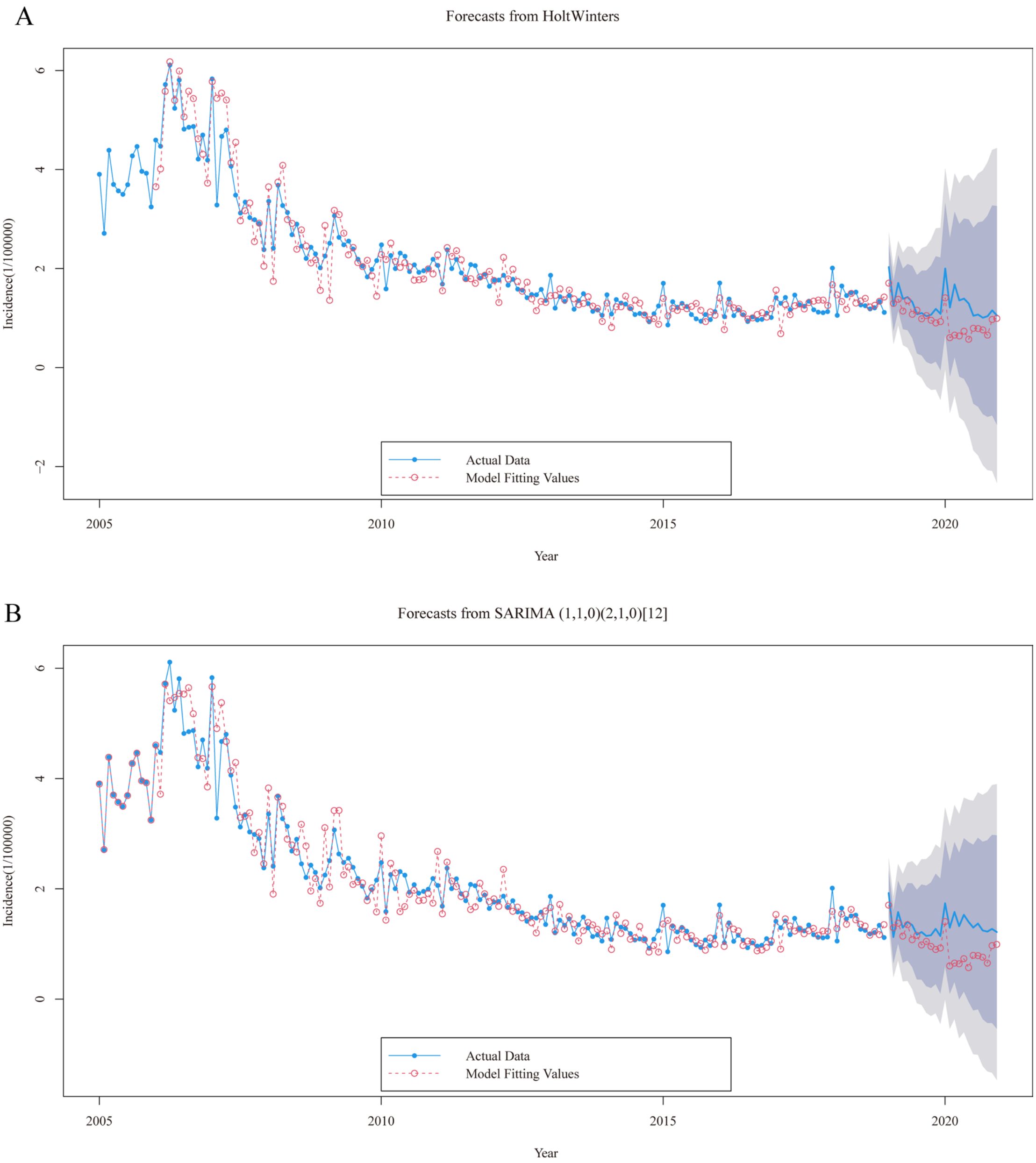

Fig. 1. Forecast plots for Holt-Winters (**A**) and SARIMA (**B**) models. The deep shaded regions indicate 80% confidence intervals, the light shaded regions indicate 95% confidence intervals.

Non-linear interaction of pollutions

From Table S4, significant interaction effects of pollutants with $PM_{2.5}$-PM_{10} and $PM_{2.5}$-CO are observed at lag periods of 5–6 months, respectively. Specifically, the interaction effect of $PM_{2.5}$-PM_{10} is better fitted at a lag of 5 month, while the interaction effect of $PM_{2.5}$-CO shows better fit at a lag of 6 months. Figure 2 illustrates fitting effect plots, revealing that the risk of viral hepatitis onset is elevated at lower levels of $PM_{2.5}$ and PM_{10} (Fig. 2A and B), while high levels of $PM_{2.5}$ (100–120 μg/m^3) and CO (Fig. 2C and D) correspond to increased onset risk. Additionally, as depicted in the fitting curves of Fig. S3, the dose–response relationships of SO_2 and NO_2 with viral hepatitis onset become progressively clearer with increasing lag months. At lag 6 month, NO_2 achieves its maximum risk effect at the level of 30–40 μg/m^3.

Discussion

The incidence of viral hepatitis in Beijing Municipality exhibited an overall decreasing trend from 2005 to 2020, primarily attributed to widespread hepatitis vaccination and standardized antiviral treatments in China. These advancements have significantly reduced new cases among patients[17]. However, despite these preventive measures, factors such as improved quality of life and various environmental influences have exacerbated the progression of hepatitis, leading to increased incidences of cirrhosis and liver cancer. Furthermore, the chronic nature of viral hepatitis, combined with limited effective prevention and treatment options, has contributed to a slight

Model	Series		Parameters	Training set			Test set		
				RMSE	R^2	MAE	RMSE	R^2	MAE
GPR (rbfdot)		Total cases	sigma = 0.476	0.163	0.582	0.120	0.383	0.002	0.325
	Classification	HAV	sigma = 0.476	0.022	0.574	0.016	0.029	0.001	0.025
		HBV	sigma = 0.476	0.117	0.517	0.082	0.226	0.033	0.195
		HCV	sigma = 0.476	0.048	0.647	0.034	0.122	0.087	0.106
		HEV	sigma = 0.476	0.031	0.400	0.024	0.050	0.080	0.042
		Unclassified hepatitis	sigma = 0.476	0.005	0.638	0.004	0.007	0.089	0.007
	Age	0–14 years	sigma = 0.476	0.003	0.412	0.003	0.004	0.000	0.004
		15–34 years	sigma = 0.476	0.055	0.518	0.042	0.119	0.000	0.105
		35–64 years	sigma = 0.476	0.097	0.608	0.071	0.217	0.000	0.185
		65- years	sigma = 0.476	0.034	0.564	0.027	0.067	0.024	0.056
GPR (polydot)		Total cases	degree = 1, scale = 1, offset = 1	0.193	0.223	0.142	0.373	0.060	0.321
	Classification	HAV	degree = 1, scale = 1, offset = 1	0.030	0.044	0.020	0.030	0.021	0.026
		HBV	degree = 1, scale = 1, offset = 1	0.137	0.174	0.097	0.222	0.003	0.194
		HCV	degree = 1, scale = 1, offset = 1	0.062	0.218	0.045	0.118	0.174	0.101
		HEV	degree = 1, scale = 1, offset = 1	0.033	0.201	0.027	0.052	0.043	0.043
		Unclassified hepatitis	degree = 1, scale = 1, offset = 1	0.007	0.074	0.005	0.008	0.000	0.007
	Age	0–14 years	degree = 1, scale = 1, offset = 1	0.004	0.115	0.003	0.004	0.009	0.004
		15–34 years	degree = 1, scale = 1, offset = 1	0.065	0.132	0.051	0.117	0.039	0.102
		35–64 years	degree = 1, scale = 1, offset = 1	0.115	0.255	0.085	0.210	0.072	0.180
		65- years	degree = 1, scale = 1, offset = 1	0.043	0.133	0.035	0.065	0.043	0.054
GPR (laplacedot)		Total cases	sigma = 0.476	0.154	0.765	0.114	0.374	0.037	0.319
	Classification	HAV	sigma = 0.476	0.021	0.722	0.014	0.029	0.001	0.025
		HBV	sigma = 0.476	0.109	0.763	0.077	0.218	0.014	0.188
		HCV	sigma = 0.476	0.047	0.741	0.033	0.121	0.148	0.106
		HEV	sigma = 0.476	0.028	0.651	0.023	0.050	0.067	0.042
		Unclassified hepatitis	sigma = 0.476	0.005	0.794	0.004	0.007	0.052	0.007
	Age	0–14 years	sigma = 0.476	0.003	0.679	0.002	0.004	0.002	0.004
		15–34 years	sigma = 0.476	0.051	0.769	0.041	0.117	0.000	0.105
		35–64 years	sigma = 0.476	0.092	0.770	0.068	0.210	0.045	0.180
		65- years	sigma = 0.476	0.033	0.729	0.026	0.065	0.069	0.055
GPR (besseldot)		Total cases	sigma = 1, order = 1, degree = 1	0.192	0.276	0.142	0.366	0.151	0.307
	Classification	HAV	sigma = 1, order = 1, degree = 1	0.028	0.197	0.019	0.031	0.000	0.027
		HBV	sigma = 1, order = 1, degree = 1	0.135	0.248	0.097	0.211	0.022	0.184
		HCV	sigma = 1, order = 1, degree = 1	0.058	0.338	0.041	0.120	0.202	0.106
		HEV	sigma = 1, order = 1, degree = 1	0.033	0.235	0.027	0.049	0.085	0.042
		Unclassified hepatitis	sigma = 1, order = 1, degree = 1	0.006	0.361	0.005	0.007	0.010	0.007
	Age	0–14 years	sigma = 1, order = 1, degree = 1	0.004	0.286	0.003	0.005	0.000	0.004
		15–34 years	sigma = 1, order = 1, degree = 1	0.065	0.213	0.051	0.117	0.046	0.105
		35–64 years	sigma = 1, order = 1, degree = 1	0.115	0.306	0.084	0.204	0.150	0.171
		65- years	sigma = 1, order = 1, degree = 1	0.041	0.282	0.032	0.063	0.133	0.053

Table 2. Comparison of the prediction results with different kernal of gaussian distribution regression (GPR) models.

rise in long-term mortality rates. The primary types of hepatitis in this region are HBV (Hepatitis B Virus) and HCV (Hepatitis C Virus). HBV transmission, particularly from mother to child, has historically been prevalent in China due to inadequate medical hygiene practices in the past. In contrast, HCV, which often presents with subtle symptoms and is not typically part of routine health screenings, has also contributed to its spread. Our study identified distinct seasonal patterns, with spring and summer showing higher incidence rates. The age group most susceptible to infection was predominantly 35–64 years old, consistent with findings from previous research[18]. This age distribution reflects the prolonged duration of hepatitis infections, with older individuals typically experiencing longer periods of infection.

Establishing robust statistical models is essential for predicting the occurrence trends of infectious diseases. Commonly utilized in time series analysis are models like Holt-Winters and ARIMA, each offering distinct advantages for predictive accuracy and practical application. In the context of viral hepatitis prediction, this study compared the Holt-Winters model with SARIMA and found that the former generally outperformed the latter. This superiority can be attributed to challenges in determining SARIMA parameters and the potential for overfitting due to complex calculations, leading to less stable predictions. The Holt-Winters model proves

				Training set			Test set		
Model	Series		Parameters	RMSE	R^2	MAE	RMSE	R^2	MAE
SVM (Linear)		Total cases	cost = 0.001, gamma = 0.2	0.221	0.125	0.167	0.354	0.196	0.306
	Classification	HAV	cost = 0.1, gamma = 0.2	0.030	0.148	0.019	0.022	0.018	0.019
		HBV	cost = 0.001, gamma = 0.2	0.154	0.168	0.108	0.200	0.215	0.168
		HCV	cost = 1, gamma = 0.2	0.056	0.364	0.037	0.107	0.004	0.090
		HEV	cost = 0.001, gamma = 0.2	0.037	0.089	0.031	0.052	0.175	0.043
		Unclassified hepatitis	cost = 0.001, gamma = 0.2	0.007	0.120	0.006	0.007	0.001	0.007
	Age	0–14 years	cost = 5, gamma = 0.2	0.003	0.366	0.002	0.004	0.001	0.003
		15–34 years	cost = 0.001, gamma = 0.2	0.070	0.173	0.056	0.110	0.065	0.098
		35–64 years	cost = 1, gamma = 0.2	0.113	0.313	0.073	0.214	0.061	0.183
		65- years	cost = 0.01, gamma = 0.2	0.047	0.099	0.037	0.064	0.132	0.055
SVM (Polynomial)		Total cases	degree = 3, cost = 0.5, gamma = 0.2	0.200	0.231	0.135	0.374	0.182	0.325
	Classification	HAV	degree = 3, cost = 0.1, gamma = 0.2	0.030	0.148	0.019	0.022	0.018	0.019
		HBV	degree = 3, cost = 0.5, gamma = 0.2	0.141	0.221	0.090	0.232	0.303	0.201
		HCV	degree = 3, cost = 1, gamma = 0.2	0.056	0.364	0.037	0.107	0.004	0.090
		HEV	degree = 3, cost = 0.1, gamma = 0.2	0.036	0.164	0.029	0.050	0.065	0.041
		Unclassified hepatitis	degree = 3, cost = 0.1, gamma = 0.2	0.007	0.143	0.005	0.007	0.001	0.006
	Age	0–14 years	degree = 3, cost = 3, gamma = 0.2	0.003	0.368	0.002	0.004	0.001	0.003
		15–34 years	degree = 3, cost = 0.1, gamma = 0.2	0.068	0.186	0.054	0.112	0.023	0.100
		35–64 years	degree = 3, cost = 1, gamma = 0.2	0.113	0.313	0.073	0.214	0.061	0.183
		65- years	degree = 3, cost = 0.1, gamma = 0.2	0.045	0.175	0.035	0.065	0.058	0.055
SVM (Radial)		Total cases	cost = 1, gamma = 1	0.150	0.594	0.085	0.393	0.001	0.349
	Classification	HAV	cost = 1, gamma = 4	0.020	0.680	0.007	0.025	0.012	0.022
		HBV	cost = 1, gamma = 0.1	0.141	0.191	0.089	0.234	0.264	0.203
		HCV	cost = 1, gamma = 0.5	0.048	0.581	0.030	0.105	0.036	0.090
		HEV	cost = 1, gamma = 0.1	0.033	0.237	0.026	0.048	0.023	0.039
		Unclassified hepatitis	cost = 1, gamma = 4	0.003	0.882	0.002	0.007	0.000	0.007
	Age	0–14 years	cost = 1, gamma = 0.5	0.004	0.358	0.002	0.004	0.009	0.003
		15–34 years	cost = 1, gamma = 0.1	0.065	0.173	0.048	0.126	0.001	0.114
		35–64 years	cost = 1, gamma = 1	0.089	0.645	0.052	0.217	0.001	0.188
		65- years	cost = 1, gamma = 0.1	0.042	0.232	0.032	0.066	0.010	0.056
SVM (Sigmoid)		Total cases	coef0 = 0.1, gamma = 1	0.150	0.594	0.085	0.393	0.001	0.349
	Classification	HAV	coef0 = 0.1, gamma = 4	0.020	0.680	0.007	0.025	0.012	0.022
		HBV	coef0 = 0.1, gamma = 0.1	0.141	0.191	0.089	0.234	0.264	0.203
		HCV	coef0 = 0.1, gamma = 0.5	0.048	0.581	0.030	0.105	0.036	0.090
		HEV	coef0 = 0.1, gamma = 0.1	0.033	0.237	0.026	0.048	0.023	0.039
		Unclassified hepatitis	coef0 = 0.1, gamma = 4	0.003	0.882	0.002	0.007	0.000	0.007
	Age	0–14 years	coef0 = 0.1, gamma = 0.5	0.004	0.358	0.002	0.004	0.009	0.003
		15–34 years	coef0 = 0.1, gamma = 0.1	0.065	0.173	0.048	0.126	0.001	0.114
		35–64 years	coef0 = 0.1, gamma = 1	0.089	0.645	0.052	0.217	0.001	0.188
		65- years	coef0 = 0.1, gamma = 0.1	0.042	0.232	0.032	0.066	0.010	0.056

Table 3. Comparison of the prediction results with different kernal of support vector machines (SVM) models.

effective in capturing epidemiological patterns of hepatitis onset due to its computational simplicity and high predictive accuracy[19]. Furthermore, this study employs machine learning-based methods to predict hepatitis onset risks associated with pollutant levels. Evaluation across different hepatitis types and age groups consistently shows superior predictive performance for primary hepatitis types and highly susceptible populations, aligning with epidemiological insights. This underscores that individuals in sensitive demographics are more vulnerable to environmental pollutants, influencing hepatitis susceptibility.

Series	Mixtures	Single pollution regression survey-weighted		Multiple pollution regression survey-weighted	
		β_{WQS} (95%CI)	p-Value	β_{WQS} (95%CI)	p-Value
Total	SO_2	0.0074 (– 0.0091, 0.0239)	0.3829	**0.0887 (0.0118, 0.1657)**	**0.0284***
	$PM_{2.5}$	**– 0.0050 (– 0.0089, – 0.0013)**	**0.0116***		
HAV	SO_2	– 0.0004 (– 0.0016, 0.0007)	0.4625	**0.0099 (0.0021, 0.0177)**	**0.016***
	CO	**– 0.0355 (– 0.0695, – 0.0016)**	**0.0461***		
	NO_2	– 0.0002 (– 0.0021, 0.0016)	0.8017		
HBV	SO_2	0.0013 (– 0.0038, 0.0065)	0.617	0.0112 (– 0.0197, 0.0421)	0.48
	$PM_{2.5}$	– 0.0021(– 0.0049, 0.0005)	0.1222		
HCV	SO_2	**0.0022 (0.0004, 0.0040)**	**0.02197***	**0.0342 (0.0210, 0.0474)**	**6.34E-06*****
	$PM_{2.5}$	**– 0.0013 (– 0.0024, – 0.0002)**	**0.02201***		
HEV	CO	– 0.0028 (– 0.0515, 0.0460)	0.9117	**0.0115 (0.0015, 0.1556)**	**0.0286***
	SO_2	0.0005 (– 0.0008, 0.0019)	0.425		
	$PM_{2.5}$	**– 0.0014 (-0.0026, – 0.0002)**	**0.0229***		
	PM_{10}	**0.0009 (0.0002, 0.0016)**	**0.0117***		
Unclassified hepatitis	$PM_{2.5}$	– 0.0002 (– 0.0004, 0.0001)	0.156	0.0017 (– 0.0001, 0.0035)	0.064561
	PM_{10}	0.0001 (– 2.47E–05, 0.0003)	0.111		
	CO	– 0.0002 (– 0.0103, 0.0097)	0.9565		
	NO_2	0.0001 (– 0.0004, 0.0006)	0.6488		
0–14 years	SO_2	– 0.0001 (– 0.0002, 0.0001)	0.3878	0.0001 (– 0.0010, 0.0013)	0.830809
	PM_{10}	– 3.84E–06 (– 0.0001, 0.0001)	0.9379		
	NO_2	**– 0.0002 (– 0.0004, – 5.14E– 06)**	**0.0499***		
15–34 years	SO_2	0.0001 (– 0.0025, 0.0027)	0.96	**0.0232 (0.0066, 0.1556)**	**0.0086****
	$PM_{2.5}$	– 0.0010 (– 0.0024, 0.0003)	0.1297		
35–64 years	SO_2	**0.0043 (0.0005, 0.0080)**	**0.03024***	**0.0453 (0.0153, 0.1556)**	**0.00473****
	$PM_{2.5}$	**– 0.0032 (– 0.0054, – 0.0009)**	**0.00793****		
65- years	SO_2	0.0001 (– 0.0017, 0.0019)	0.897	**0.0127 (0.0009, 0.1556)**	**0.0408***
	$PM_{2.5}$	– 0.0002 (– 0.0009, 0.0006)	0.652		

Table 4. Comparison of results from the survey-weighted single pollution analyses and WQS regression of the matrix specific pollutions mixtures for the viral hepatitis. The parameter estimate (β) is reported in bold for significant single pollution or WQS mixture effects. The components with the highest weights are reported for mixtures with significant effects. Bold font indicates statistical significance at the 0.05 level. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$.

Different types of viral hepatitis primarily spread through gastrointestinal and bloodborne routes. HAV and HEV, for instance, mainly transmit through the gastrointestinal tract, with transmission influenced by pollutants such as PM_{10} and CO. This can be linked to increasing industrialization and declining environmental awareness. Higher levels of airborne particulate matter and vehicle emissions exacerbate environmental pollution, thereby enhancing transmission through the gastrointestinal route. Other types of viral hepatitis primarily transmit through blood and bodily fluids, affected notably by pollutants like SO_2 and $PM_{2.5}$. Epidemiological studies have shown an association between $PM_{2.5}$ levels and liver fibrosis[20]. Animal research indicates that air pollution can activate Kupffer cells, trigger endoplasmic reticulum stress responses, induce cytokine production, and promote collagen deposition, thereby exacerbating fibrosis progression[21]. This suggests environmental pollutants can impact hepatic metabolism through the bloodstream route. Furthermore, this study identifies SO_2 and CO as significant pollutants influencing the onset and mortality of viral hepatitis. CO, due to its high affinity for hemoglobin binding in the bloodstream, poses a notable threat to the progression and mortality of hepatitis. These findings underscore the importance highlighted in China's infectious disease planning of addressing hepatitis transmitted through the bloodstream route.

Current literature on infectious disease prediction and pollutant impacts often focuses on single methodologies and specific effects. This study, however, employed diverse time-series methods to forecast and analyze the interactive effects of viral hepatitis, revealing significant month-to-month prediction intervals marked by considerable fluctuations. These findings underscore the challenge of capturing the inherent volatility in viral hepatitis data using conventional models. Moreover, regional constraints within the study area limited the generalizability of findings across different types of hepatitis affected by pollutants. Future research endeavors are encouraged to validate these macroscopic epidemiological insights at a microscopic level, utilizing animal models to elucidate underlying physiological mechanisms.

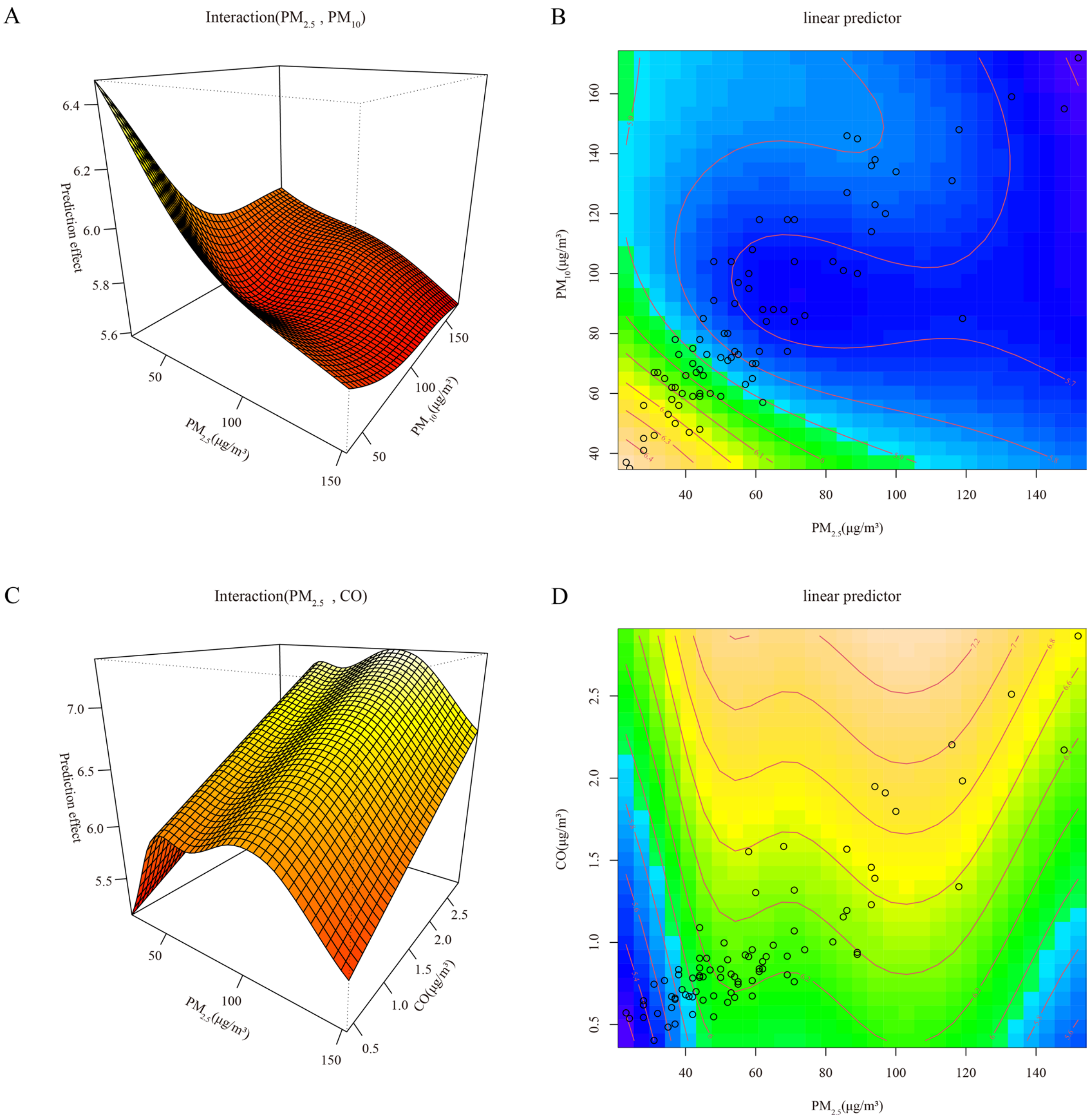

Fig. 2. The fitting interactions of the association among pollutants and viral hepatitis cases in Beijing, 2014–2020 based on the generalized additive model (GAM), with lagging of 5 (**A**, **B**) and 6 (**C**, **D**) months.

Conclusion

The Holt-Winters model outperformed SARIMA in predicting viral hepatitis incidence. SVM and GPR models utilizing pollutant data showed potential for enhanced prediction accuracy. Patients with HAV and HEV were primarily impacted by PM_{10} and CO, while SO_2 and $PM_{2.5}$ affected other types. The 35–64 age group exhibited higher susceptibility. Long-term exposure to mixed pollutants influenced hepatitis development with a lag of 5–6 months, emphasizing the need for sustained pollutant monitoring for effective public health strategies.

Data availability

The data that support the findings of this study are available on request from the National Public Health Data Centre of China (https://www.phsciencedata.cn/) and the National Oceanic and Atmospheric Administration (NOAA) (https://www.noaa.gov/).

Received: 19 April 2024; Accepted: 3 September 2024
Published online: 10 September 2024

References

1. Liou, J. W., Mani, H. & Yen, J. H. Viral hepatitis, cholesterol metabolism, and cholesterol-lowering natural compounds. *Int. J. Mol. Sci.* **23**, 7 (2022).
2. Pisano, M. B. *et al.* Viral hepatitis update: Progress and perspectives. *World J. Gastroenterol.* **27**(26), 4018–4044 (2021).
3. Martin, N. A. The discovery of viral hepatitis: A military perspective. *J. R. Army Med. Corps* **149**(2), 121–124 (2003).

4. Stanaway, J. D. *et al.* The global burden of viral hepatitis from 1990 to 2013: Findings from the Global Burden of Disease Study 2013. *Lancet* **388**(10049), 1081–1088 (2016).
5. Schweitzer, A., Horn, J., Mikolajczyk, R. T., Krause, G. & Ott, J. J. Estimations of worldwide prevalence of chronic hepatitis B virus infection: A systematic review of data published between 1965 and 2013. *Lancet* **386**(10003), 1546–1555 (2015).
6. Yue, T. *et al.* Trends in the disease burden of HBV and HCV infection in China from 1990–2019. *Int. J. Infect. Dis.* **122**, 476–485 (2022).
7. 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.* **15**(1), 28–31 (2015).
8. Gullón, P., Varela, C., Martínez, E. V. & Gómez-Barroso, D. Association between meteorological factors and hepatitis A in Spain 2010–2014. *Environ. Int.* **102**, 230–235 (2017).
9. Jang, T. Y., Ho, C. C., Wu, C. D., Dai, C. Y. & Chen, P. C. Air pollution as a potential risk factor for hepatocellular carcinoma in Taiwanese patients after adjusting for chronic viral hepatitis. *J. Chin. Med. Assoc.* **87**(3), 287–291 (2024).
10. Wang, S., Wei, F., Li, H., Wang, Z. & Wei, P. Comparison of SARIMA model and Holt-Winters model in predicting the incidence of Sjögren's syndrome. *Int. J. Rheum. Dis.* **25**(11), 1263–1269 (2022).
11. Nath, P., Saha, P., Middya, A. I. & Roy, S. Long-term time-series pollution forecast using statistical and deep learning methods. *Neural Comput. Appl.* **33**(19), 12551–12570 (2021).
12. Zhu, X. *et al.* Prediction study of electric energy production in important power production base, China. *Sci. Rep.* **12**(1), 21472 (2022).
13. Chen, Y., Hou, W. & Dong, J. Time series analyses based on the joint lagged effect analysis of pollution and meteorological factors of hemorrhagic fever with renal syndrome and the construction of prediction model. *PLoS Negl. Trop. Dis.* **17**(7), e0010806 (2023).
14. Cole, J. H. *et al.* Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage* **163**, 115–124 (2017).
15. Pisner, D. A. & Schnyer, D. M. J. M. L. *Support vector machine—ScienceDirect* 101–121 (Elsevier, 2020).
16. Xu, J. *et al.* Associations of metal exposure with hyperuricemia and gout in general adults. *Front. Endocrinol.* **13**, 1052784 (2022).
17. Liang, X. *et al.* Epidemiological serosurvey of hepatitis B in China–declining HBV prevalence due to hepatitis B vaccination. *Vaccine* **27**(47), 6550–6557 (2009).
18. Bai, H., Liu, H., Chen, X., Xu, C. & Dou, X. Influence of age and HBeAg status on the correlation between HBV DNA and hepatic inflammation and fibrosis in chronic hepatitis B patients. *Digest. Dis. Sci.* **58**(5), 1355–1362 (2013).
19. Zhou, Y. *et al.* Trend of the tuberculous pleurisy notification rate in Eastern China During 2017–2021: Spatiotemporal analysis. *JMIR Public Health Surveill.* **9**, e49859 (2023).
20. Jang, T. Y. *et al.* Air pollution associate with advanced hepatic fibrosis among patients with chronic liver disease. *Kaohsiung J. Med. Sci.* **40**(3), 304–314 (2024).
21. Zheng, Z. *et al.* Exposure to fine airborne particulate matters induces hepatic fibrosis in murine models. *J. Hepatol.* **63**(6), 1397–1404 (2015).

Acknowledgements

The disease data was publicly supported from the National Public Health Data Centre of China (https://www.phsciencedata.cn/). Pollutions information was publicly from the National Oceanic and Atmospheric Administration (NOAA) (https://www.noaa.gov/) including CO, NO2, O3 etc.

Author contributions

SF P: Software, Conceptual, Methodology, Formal analysis, Investigation, Resources, Writing-original draft, Writing-review & editing. L Y: Software, Conceptual, Methodology, Formal analysis, Investigation, Writing-original draft, Writing-review & editing. HX G: Conceptualization, Methodology, Formal analysis, Writing-original draft, Writing-review & editing, Funding acquisition, Supervision. YZ L: Methodology, Software, Writing—original draft, Visualization. EH D: Methodology, Software, Writing—original draft, Visualization. FM F: Conceptualization, Methodology, Formal analysis, Writing-review & editing, Funding acquisition, Supervision. JH L: Conceptualization, Methodology, Formal analysis, Writing-review & editing, Funding acquisition, Supervision. All authors had full access to the data, contributed to the study, approved the final version for publication, and take responsibility for its accuracy and integrity. All authors read and approved the final manuscript.

Funding

This work was supported by the Medical Research Program of Hebei Province (20231696) and the National Natural Science Foundation of China (81670525).

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Viral hepatitis is a Class B infectious disease under China's Infectious Disease Prevention and Control Law, and each case reported by a medical institution is reported through the direct reporting system of the infectious disease network and requires epidemiological investigation and surveillance testing to further clarify the source of the virus and infection. Specimens are first tested by the laboratories of medical institutions, of which positive specimens are reviewed by disease prevention and control institutions and the results are fed back to the sending units. Therefore, this study received ethical approval from the China Center for Disease Control and Prevention. Since the disease under investigation, viral hepatitis, is a statutory infectious disease subject to national statutory monitoring each year, informed consent is not required. For confidentiality reasons, all viral hepatitis data were analyzed anonymously.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-72047-1.

Correspondence and requests for materials should be addressed to F.F. or J.L.

Reprints and permissions information is available at www.nature.com/reprints.