Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Environmental Mercury and Persistent Organic Pollutants as Key Drivers of Fecal-Oral Viral Hepatitis: A Time-Series and Machine Learning Analysis

Weiming Hou

hwm100908@163.com

Department of Medical Engineering, Air Force Medical Center, PLA, Air Force Medical University, Beijing, 100142, Peoples' Republic of China.

Research Article

Keywords: Viral hepatitis, Gaseous oxidized mercury, Persistent Organic Pollutants, Time series analysis, Weighted quantile sum

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11062734/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Hepatitis A virus (HAV) and hepatitis E virus (HEV) are both acutely transmitted via the fecal-oral route and pose a significant global burden of infection. Existing disease predictions often do not account for environmental chemical pollutants, and there is insufficient epidemiological evidence linking mercury and persistent organic pollutants (POPs) to hepatitis incidence and mortality.

Methods

This study utilized hepatitis surveillance data from China between 2005 and 2020 along with pollutant emission data from the EDGAR database. A SARIMA model was employed for univariate time series forecasting, combined with a BP neural network for multivariate predictions. Key toxic factors were identified using LASSO regression, and WQS regression was applied to analyze mixed exposure effects stratified by age. Furthermore, a random forest model was used to explore sensitive factors and interactions during different dominant periods of hepatitis epidemics.

Results

The SARIMA monthly model demonstrated superior predictive performance for HAV and HEV incidence, while the quarterly model was found to be more suitable for mortality rate predictions; overall, the predictive performance for HAV was slightly better than that for HEV. Gas-phase oxidized mercury (GOM), particulate-bound mercury (PBM), polycyclic aromatic hydrocarbons (PAHs), and polychlorinated biphenyls (PCBs) were identified as core influencing factors. There was significant heterogeneity in key toxicants and risk effects across different age groups, with those aged 60 and above being more sensitive to particulate-bound mercury exposure. The shift in dominant hepatitis epidemic types significantly altered environmental driving factors: the HAV-dominant period was characterized mainly by GOM, PCB, and HCB, while the HEV-dominant period was characterized by GEM, PCDD/F, and HCB. The dual-dominant period exhibited the most pronounced features of co-exposure to multiple toxicants and yielded the best model fitting results.

Conclusions

Environmental mercury and POPs are significantly associated with the risk of HAV and HEV infections, with notable age differences and heterogeneous effects during different epidemic periods. Integrating environmental indicators into machine learning and time series models can effectively enhance the predictive accuracy of hepatitis incidence and mortality rates. The findings provide a scientific basis for

environmental health risk prevention and targeted interventions for hepatitis across different populations and stages.

1. Introduction

Hepatitis A virus (HAV) is a globally transmitted disease that spreads via the fecal-oral route[1], including the ingestion of food and water contaminated with HAV, as well as through person-to-person contact, such as exposure to household items of infected individuals, sexual transmission, and international travel[2–4]. It is estimated that there are approximately 1.4 million new cases of HAV infection worldwide each year. Since the development of a safe and effective hepatitis A vaccine in the early 1990s, understanding the epidemiology of hepatitis A has become increasingly important for making informed decisions regarding appropriate preventive measures[5]. Hepatitis E virus (HEV), belonging to the family Hepeviridae and the genus Hepevirus, is an enveloped virus with single-stranded positive-sense RNA[6]. Acute HEV infection rarely presents with clinical symptoms in children[7]. Infections caused by the non-enveloped viruses HAV and HEV are typically asymptomatic or mild. When symptoms do occur, they present in an acute form, requiring only symptomatic treatment without specific antiviral therapy[8, 9]. Due to its acute nature and mild symptoms, the impact on healthcare systems is minimal. The predictability and complexity of human infections caused by these viruses remain uncertain.

Time series analysis of infection data through specific models can enhance preventive systems and predict future values based on previously observed data[10]. These models include Seasonal Autoregressive Integrated Moving Average (SARIMA), neural network models, exponential smoothing, gray wave models, and others[11–13]. Notably, the SARIMA model excels in accurately identifying the seasonality and trends of infectious diseases. For instance, Chen Ye et al. utilized the SARIMA method to predict the incidence of hemorrhagic fever[14]. Yicheng Liang et al. applied ARIMA to forecast the future disease burden for the year 2050[15]. Although previous studies have extensively explored disease prediction, researchers often overlook the effects of different disease subtypes and there exists a certain gap in mortality rate predictions. Additionally, there is a lack of consideration of the potential impact of chemical pollutants in predictive studies. There is a degree of correlation between environmental indicators and disease incidence[16, 17]. Therefore, incorporating environmental indicators into infectious disease prediction models is expected to enhance the accuracy of predictions to some extent.

Mercury is considered one of the major harmful environmental pollutants due to its widespread circulation in ecosystems and its inability to be eliminated from the environment[18]. The toxicity of mercury and its neurodegenerative effects have been confirmed by numerous studies[19]. Research has found that prenatal mercury exposure is associated with elevated liver enzyme levels in children, and this association may be mediated by changes in DNA methylation and gene expression[20]. Persistent organic pollutants (POPs) are a class of environmental chemicals that are widely present, highly stable, and capable of bioaccumulation, including polychlorinated biphenyls (PCBs), dioxins, furans, and various organochlorine pesticides. These substances can enter the human body through food, air, and drinking water, accumulating in adipose tissue over time. Experimental studies suggest that POPs can induce

fatty liver degeneration and inflammatory responses. However, there is still limited epidemiological evidence regarding these environmental toxicants in populations affected by hepatitis, and the specific research findings are inconsistent[21–23]. There is an urgent need for population-based epidemiological studies.

Several studies have shown an increasingly apparent positive and negative correlation between chemical pollution and the incidence of specific types of hepatitis. However, the effects of chemical pollution on different hepatitis subtypes and during various dominant periods may vary, although these differences have not been widely assessed. This study aims to investigate the impact of mercury-containing gases and persistent organic pollutants (POPs) on the incidence and mortality of hepatitis A and E in mainland China, as well as disease prediction. To ensure the monitoring of dual outcomes (incidence and mortality) for HAV and HEV and the assessment of the actual effects of toxic exposure, we employed time series analysis and mixture analysis methods. Additionally, to fairly evaluate the actual impact of HAV and HEV alongside toxic exposures, we conducted subgroup analyses based on age and dominant periods.

2. Materials and Methods

2.1 Data collection

Data on reported cases, incidence rates, mortality cases, and mortality rates of fecal-oral transmitted viral hepatitis (including HAV and HEV) in China from 2005 to 2020 were obtained from the Public Health Science Data Center website (https://www.phsciencedata.cn/). This website collects incidence data for various types of viral hepatitis. All cases were diagnosed according to the diagnostic criteria set forth in the viral hepatitis prevention and treatment program issued by the Ministry of Health of the People's Republic of China. The study calculated the mortality rate by using reported case and death case data.

Data on different mercury species (Gaseous Elemental Mercury (GEM), Gaseous Oxidized Mercury (GOM), and Particle-Bound Mercury (PBM)) and Unintentional Persistent Organic Pollutants (UPOPs) (including Polycyclic Aromatic Hydrocarbons (PAHs): benzo[a]pyrene (BaP), benzo[b]fluoranthene (BbF), benzo[k]fluoranthene (BkF), and Indeno[1,2,3-cd]pyrene (IcdP); Dioxins and Furans (PCDD/Fs); Hexachlorobenzene (HCB); and Polychlorinated Biphenyls (PCBs)) from 2005 to 2020 were sourced from the Global Atmospheric Research Emissions Database (EDGAR) (https://edgar.jrc.ec.europa.eu/). This website collects emission data for various pollutants, with values reported in gigagrams (Gg).

2.2 Prediction models

2.2.1 Single factor prediction- Seasonal Autoregressive Integrated Moving Average (SARIMA) model

This study utilized the SARIMA model to forecast the incidence and mortality rates of HAV and HEV on a quarterly and monthly time series basis. The time series data from 2005 to 2018 were used to establish

the training set, while the time series data from 2019 to 2020 were used to create the testing set.

The SARIMA model consists of four components: Autoregressive (AR), Integrated (I), Moving Average (MA), and Seasonal effects (S). SARIMA can be categorized into multiplicative or additive models based on the relationship between seasonal effects and the other components. In the multiplicative model, the expression is SARIMA(p,d,q) x (P,D,Q)s, where p and P represent the orders of the regular and seasonal autoregressive terms, respectively; d and D denote the orders of differencing for the regular and seasonal components; q and Q correspond to the orders of the regular and seasonal moving average terms, respectively; and S represents the length of the seasonal period[24]. The parameters are typically estimated based on the orders of the ACF and PACF plots, which has certain limitations. In this study, we first determined the initial orders using the auto.arima() function, and then obtained the coefficients and corresponding standard errors (SE) from the automatic order selection. A coefficient greater than 2 SE indicates significant relevance ($P < 0.05$). We also calculated and compared the AIC values, ultimately selecting the appropriate order for modeling based on both significance and the minimum AIC value. The Augmented Dickey-Fuller (ADF) test was used to examine the stationarity of the time series. A p-value < 0.05 indicated a stationary series. The Box-Ljung test was applied to check whether the model residuals were white noise. A p-value > 0.05 suggested no significant autocorrelation in the residuals, indicating adequate model fitting. The model was evaluated based on the RMSE, MAE, and MAPE metrics, with smaller values indicating better performance.

2.2.2 Multiple factor prediction-Back Propagation (BP) neural network model

Before applying the multivariate BP neural network algorithm for prediction, the study utilized Spearman correlation analysis to filter out chemical factors with strong correlations. The BP neural network primarily relies on the error backpropagation algorithm, which iteratively adjusts weights and biases to minimize the prediction error at the output layer. The BP neural network model consists of an input layer, hidden layers, and an output layer[25]. The input layer receives the emission data of toxic chemicals, which is then transmitted and adjusted multiple times through the hidden layers, ultimately producing the analysis results for hepatitis incidence and mortality rates. The number of neurons in the hidden layers is crucial for the performance of the neural network. In the initial phase of this study, after testing and reviewing the literature, three hidden layers with 5, 5, and 5 neurons were added after the input layer for the incidence rate of HAV. For the mortality rate of HAV, two hidden layers with 4 and 4 neurons were added. For the incidence rate of HEV, three hidden layers with 6, 6, and 6 neurons were added. For the mortality rate of HEV, a single hidden layer with 5 neurons was included after the input layer.

The training method employed was Adaptive Gradient Descent, with two types of activation functions set for modeling. In Model 1, the activation function for the neurons in the hidden layers was Sigmoid, while the activation function for the neurons in the output layer was purelin. In Model 2, the activation function for the neurons in the hidden layers was also Sigmoid, and the activation function for the output layer

neurons was tansig[16]. The models were evaluated based on RMSE, MAE, and MAPE metrics, with smaller values indicating better performance.

2.3 Features selection

To efficiently select toxic factor variables and address multicollinearity, the study employed regularization techniques to reduce the correlations among toxic factors. Among these techniques, the Least Absolute Shrinkage and Selection Operator (LASSO) regression model demonstrated significant advantages. The selection groups included the overall HAV and HEV incidence populations, as well as age-specific groups for HAV and HEV incidence (including ages 0–14, 15–39, 40–59, and 60 and above). In this study, LASSO regression analysis was conducted using the glmnet package in R. By incorporating optimal performance parameters, the analysis avoided overfitting, making the model easier to interpret. LASSO automatically shrinks some coefficients to zero by penalizing the sum of the absolute coefficients, selectively retaining only the most relevant variables and reducing the impact of less correlated predictors[26]. The value of λ was chosen within the range of minimum error ± 1 standard error, resulting in fewer and more stable selected variables. The iterative analysis employed a 5-fold cross-validation method.

2.4 Stratified mixtures analysis method in aged group

This study applied weighted quantile sum (WQS) regression to investigate the risk effects of individual toxic factors and chemical mixtures on the incidence of HAV and HEV. The analysis was conducted not only for the overall population but also for different age-specific groups (including ages 0–14, 15–39, 40–59, and 60 and above). Additionally, the results were visualized to identify positive risk effect factors and to observe the dose-response relationship between the toxic mixtures and the incidence risk.

This study used the R package (gWQS) to calculate the weighted WQS index, a statistical model suitable for multivariable regression analysis of high-dimensional datasets. The WQS index represents the weighted sum of the emissions of various toxic factors, where each exposure variable is assigned a corresponding weight to estimate its relative contribution to the outcome. The index ranges from 0 to 1, reflecting the overall burden of exposure risk associated with the toxic factors. In this research, a forward WQS index was constructed through 1,000 bootstrap resampling iterations to estimate the relative contributions of each toxic factor to HAV and HEV. The dataset was randomly divided into two parts: 60% as the training set and 40% as the validation set. The training set was used to estimate the weights of the exposure factors, while the validation set was employed to test the statistical significance of the WQS index. For each bootstrap sample, a new dataset was generated by resampling with replacement from the training set, and an optimization algorithm was used to estimate the model parameters. Ultimately, the WQS index was obtained by averaging the weights from multiple bootstrap samples. After constructing the WQS index, the association between the weighted exposure mixture and the outcome was examined in the validation set. This method quantitatively estimates the relative contribution of each toxic factor to the overall exposure risk, providing a comprehensive evaluation of its cumulative impact on the outcomes[27].

2.5 Period-Dependent Associations between Environmental Exposures and Viral Hepatitis:

2.5.1 Period-Dependent Associations between Environmental Exposures and Viral Hepatitis: Sensitive Factor Identification

The study classified the incidence of HAV and HEV in China from 2005 to 2020 into three types: HAV-dominant, HEV-dominant, and dual-dominant. The classification thresholds were defined as follows: HAV-dominant type: HAV composition ratio ≥ 60% with statistically significant differences; HEV-dominant type: HEV composition ratio ≥ 60% with statistically significant differences; dual-dominant type: both types with composition ratios between 40% and 60%, with no statistically significant differences. First, the study explored the seasonal and time series characteristics under different classifications. Second, for each classification period, a random forest model was employed to display the importance ranking and significance levels of various environmental factors under different dominant periods. The model's fit was assessed using R^2 values and overall model p-values. This analysis utilized the “rfPermute” package in R to construct the random forest model, followed by significance testing of the entire model using the “A3” package, which provided R^2 values for subsequent graphical annotations[28]. The importance ranking was determined using the Increase in MSE (Mean Squared Error increase percentage) metric, which is the core indicator for measuring variable importance in random forests. A higher value indicates that the variable contributes more significantly to the model's predictive accuracy, thus having a stronger impact on hepatitis incidence.

2.5.2 Period-Dependent Associations between Environmental Exposures and Viral Hepatitis: Interaction Assessment

The study utilized the "randomForestExplainer" package in R to compute the depth distribution of each toxic feature, providing an approximate understanding of the importance of these features. Subsequently, based on the mean_min_depth, the most significant variables among the previously identified toxic factors were selected, allowing for the determination of the minimum depth information regarding the interactions between each toxic factor. This information was then visualized to identify the group of interacting factors with the smallest mean_min_depth, and further exploration was conducted to assess how these interactions impacted the prediction of HAV and HEV incidence outcomes. A heatmap was generated to display the model's predicted values for hepatitis incidence under different concentration combinations of the two core variables: red indicating high disease risk and blue indicating low disease risk. The horizontal and vertical axes represent the concentration gradients of the two variables, while the color gradient visually reflects the dose-response relationship between exposure and disease incidence.

3. Results

3.1 The analysis of SARIMA model results

According to the univariate prediction results in Table 1, for predicting the incidence of HAV and HEV, the SARIMA Month Model outperforms the SARIMA Quarter Model (with lower RMSE and MAE); the models are (3,1,1) (2,1,1) [12] and (0,1,2) (0,1,1) [12], respectively. In contrast, for predicting the mortality rates of HAV and HEV, the SARIMA Quarter Model shows greater advantages (with lower RMSE and MAE); the models are (1,0,1) (0,1,0) [12] and (1,1,2) (0,1,1) [12], respectively.

Table 1
Parameter fitting and error assessment of different prediction models for the incidence and mortality rates of HAV and HEV.

Model	Series	Factors	Best parameters	RMSE	MAE	MAPE
SARIMA Month Model	HAV Incidence	Time	(3,1,1) (2,1,1) [12]	0.0253	0.0177	7.9150
SARIMA Quarter Model		Time	(2,1,1) (1,0,0) [12]	0.0819	0.0556	7.5372
BP neural network Model 1		Toxic factors	n.neurons = c(7,5,5,5,1), hidden.layer="sigmoid", output.layer="purelin"	0.0362	0.0271	13.0239
BP neural network Model 2		Toxic factors	n.neurons = c(7,5,5,5,1), hidden.layer="sigmoid", output.layer="tansig"	0.0409	0.0320	15.8766
SARIMA Month Model	HAV Fatality	Time	(1,0,1) (1,1,0) [12]	0.0004	0.0003	Inf
SARIMA Quarter Model		Time	(1,0,1) (0,1,0) [12]	0.0003	0.0002	Inf
BP neural network Model 1		Toxic factors	n.neurons = c(4,4,4,1), hidden.layer="sigmoid", output.layer="purelin"	0.0003	0.0003	10606.0200
BP neural network Model 2		Toxic factors	n.neurons = c(4,4,4,1), hidden.layer="sigmoid", output.layer="tansig"	0.0003	0.0003	10131.5700
SARIMA Month Model	HEV Incidence	Time	(0,1,2) (0,1,1) [12]	0.0143	0.0097	6.1199
SARIMA Quarter Model		Time	(0,1,1) (1,0,1) [12]	0.0397	0.0319	7.2252
BP neural network		Toxic factors	n.neurons = c(10,6,6,6,1),	0.0314	0.0239	16.6336

Abbreviations: RMSE: Root Mean Square Error;MAE: Mean Absolute Error；MAPE: Mean Absolute Percentage Error；Inf: Infinite, indicates that the model prediction error is significant.

Model	Series	Factors	Best parameters	RMSE	MAE	MAPE
Model 1			hidden.layer="sigmoid", output.layer="purelin"			
BP neural network Model 2		Toxic factors	n.neurons = c(10,6,6,6,1), hidden.layer="sigmoid", output.layer="tansig"	0.0315	0.0240	16.5792
SARIMA Month Model	HEV Fatality	Time	(1,1,2) (0,1,1) [12]	0.0009	0.0006	Inf
SARIMA Quarter Model		Time	(1,1,2) (0,1,1) [12]	0.0005	0.0003	41.0086
BP neural network Model 1		Toxic factors	n.neurons = c(5,5,1), hidden.layer="sigmoid", output.layer="purelin"	0.0008	0.0006	10555.9500
BP neural network Model 2		Toxic factors	n.neurons = c(5,5,1), hidden.layer="sigmoid", output.layer="tansig"	0.0008	0.0006	10477.6400

Abbreviations: RMSE: Root Mean Square Error;MAE: Mean Absolute Error□MAPE: Mean Absolute Percentage Error□Inf: Infinite, indicates that the model prediction error is significant.

From the prediction data and related graphs observed in **Table S1** and Figs. 1 and S1, it can be noted that under the monthly model predictions, the overall prediction error for mortality rates is relatively low. Although the monthly error for HAV incidence in 2019 (mean error 0.0272) is higher than that for HEV (mean error − 0.0047), the monthly error for HAV incidence in 2020 (mean error 0.012) is significantly lower than that for HEV (mean error − 0.0609). For the quarterly model predictions, the overall prediction error for mortality rates is also relatively low; the quarterly error for HAV incidence in 2019 (mean error 0.0607) is higher than that for HEV (mean error − 0.0117), while the quarterly error for HAV incidence in 2020 (mean error − 0.0189) is significantly lower than that for HEV (mean error − 0.1744). In summary, the results indicate that in the SARIMA monthly and quarterly model predictions, HAV incidence tends to be over-predicted, while HEV incidence tends to be under-predicted. Additionally, the prediction performance of the SARIMA model for HAV is slightly better than that for HEV.

3.2 The analysis of BP neural network model results

The results of the Spearman correlation analysis presented in **Figure S2** indicate the following factors can be selected: For factors related to HAV incidence, the mercury species include GOM and PBM, while the UPOPs include BkF, IcdP, PCDD/F, HCB, and PCB. For factors associated with HAV fatality, the mercury species again include GOM and PBM, with UPOPs comprising HCB and PCB. In the case of HEV incidence, all included factors exhibit a significant positive correlation ($r > 0$, $P < 0.05$). Regarding factors related to HEV fatality, the mercury species consist of GOM and PBM, while the UPOPs include PCDD/F and PCB.

From the multifactor prediction graphs presented in Fig. 2, it can be observed that there are significant differences among the various models only in the HAV incidence prediction graph (Fig. 2A), while the other graphs show no notable differences (Fig. 2B**/C/D**). Specifically in Fig. 2A, Model 1 demonstrates a greater number of time points that are close to the actual values compared to Model 2, particularly evident during the period from 2010 to 2015. Subsequently, based on the model metrics calculated in Table 2, it is found that, apart from the HAV incidence prediction metrics where the RMSE, MAE, and MAPE of Model 1 are lower than those of Model 2, Model 2 outperforms Model 1 in all other predictions. Additionally, regarding the testing of the BP neural network prediction model, it is noted that for both HAV and HEV, selecting a higher number of hidden layers provides an advantage for incidence predictions, while a lower number of hidden layers can optimize mortality predictions.

Table 2
Assessment of the effects of single and mixed toxic substances on the risk of HAV infection in different age groups using Weighted Quantile Sum (WQS) regression.

Series	Mixtures	Single toxics regression survey-weighted		Multiple toxics regression survey-weighted	
		β_{WQS} (95%CI)	p-Value	β_{WQS} (95%CI)	p-Value
Total	GEM	2.92 (1.70, 4.14)	**0.0000*** **	127.88 (-272.26, 528.03)	0.53
	GOM	17.53 (0.82, 34.23)	**0.0421***		
	PBM	-192.00 (-322.08, -61.92)	**0.0046****		
	HCB	-882.73 (-1608.32, -157.14)	**0.0188***		
	PAH_BbF	0.37 (0.10, 0.65)	**0.0092****		
	PAH_BkF	1.88 (0.34, 3.41)	**0.0186***		
	PAH_IcdP	-1.88 (-3.08, -0.68)	**0.0027****		
	PCB	-191.38 (-296.05, -86.72)	**0.0005*** **		
Aged 0–14 years	GEM	1.67 (0.70, 2.64)	**0.00103****	-10.44 (-172.46, 151.57)	0.90
	GOM	9.67 (1.54, 17.79)	**0.02157***		
	PBM	-121.07 (-189.57, -52.58)	**0.00076*** **		
	PAH_BbF	0.18 (0.04, 0.33)	**0.01442***		
	PAH_BkF	1.17 (0.35, 1.98)	**0.00582****		
	PAH_IcdP	-1.08 (-1.71, -0.44)	**0.00121****		
Aged 15–39 years	GEM	1.00 (0.50, 1.49)	**0.00014*** **	143.97 (-3.26, 291.20)	0.06
	GOM	6.91 (0.35, 13.46)	**0.04131***		

Abbreviations: β_{WQS}: Weighted Quantile Sum regression coefficient; 95% CI: 95% confidence interval; GEM: Gaseous elemental mercury; GOM: Gaseous oxidized mercury; PBM: Particulate-bound mercury; HCB: Hexachlorobenzene; PAH: Polycyclic aromatic hydrocarbons; PCB: Polychlorinated biphenyls.

Series	Mixtures	Single toxics regression survey-weighted		Multiple toxics regression survey-weighted	
		β_{WQS} (95%CI)	p-Value	β_{WQS} (95%CI)	p-Value
	PBM	-77.58 (-128.34, -26.82)	**0.0034****		
	HCB	-433.07 (-716.30, -149.85)	**0.00338****		
	PAH_BbF	0.19 (0.08, 0.30)	**0.00087*****		
	PAH_BkF	0.70 (0.10, 1.30)	**0.02502***		
	PAH_IcdP	-0.81 (-1.28, -0.34)	**0.00101****		
	PCB	-75.07 (-115.97, -34.17)	**0.00049*****		
Aged 40–59 years	GEM	0.36 (0.00, 0.72)	0.055	64.49 (-38.17, 167.15)	0.22
	PCB	-41.00 (-70.00, -12.00)	**0.00659****		
Aged 60 + years	PBM	20.95 (0.63, 41.27)	**0.04577***	23.64 (-42.17, 89.45)	0.48
	PAH_BbF	-0.05 (-0.09, -0.01)	**0.02058***		
	PCB	-25.14 (-41.32, -8.95)	**0.00293****		
*** P < 0.001, ** P < 0.01, * P < 0.05					
Bold font indicates statistical significance at the 0.05 level.					
Abbreviations: β_{WQS}: Weighted Quantile Sum regression coefficient; 95% CI: 95% confidence interval; GEM: Gaseous elemental mercury; GOM: Gaseous oxidized mercury; PBM: Particulate-bound mercury; HCB: Hexachlorobenzene; PAH: Polycyclic aromatic hydrocarbons; PCB: Polychlorinated biphenyls.					

3.3 Features analysis

Using Lasso for parameter selection, the variability of these variable coefficients is illustrated in **Figures S3** and **S4**. When λ values are set at 2.22 and 157.85, a high-performance overall model for HAV and HEV with the fewest variables is obtained. The screening variables for overall HAV and HEV patients, as well as for different age group stratifications, are detailed in **Table S2**. In the HAV patient cohort, a total of 9 specific factors were identified from the union (GEM, GOM, PBM, HCB, PAH_BbF, PAH_BkF, PAH_IcdP, PCB, PCDD_F), with GEM and PAH_BbF identified as positive feature factors. In the HEV

patient cohort, a total of 6 specific factors were identified from the union (GEM, GOM, PBM, PAH_IcdP, PCB, PCDD_F), with GEM and PCB identified as positive feature factors.

3.4 Mixtures analysis

The results of WQS mixture modeling for HAV and HEV after feature selection using Lasso regression are presented in Tables 2 and 3. According to Table 2, in the overall population of HAV and in the age groups 0–14 years and 15–39 years, GEM, GOM, PAH_BbF, and PAH_BkF all show significant positive risk effects on the number of cases. The risk effect values (β) are ranked from smallest to largest as follows: PAH_BbF > PAH_BkF > GEM > GOM. In the HAV population aged 60 years and older, PBM exhibits a significant positive risk effect on the number of cases (βWQS (95%CI): 20.95 (0.63, 41.27)). Among the HAV subgroups, the overall toxic mixture shows non-significant risk effects on the number of cases except for the group aged 0–14 years. From Table 3, it is evident that, with the exception of the HEV population aged 0–14 years, PBM demonstrates a significant positive risk effect on the number of cases across all age groups. The range of risk effects (β) is [1.48, 80.60]. In the HEV populations aged 15–39 years and 60 years and older, the significance level of the risk associated with PBM is notably high ($P < 0.01$), with effects (βWQS (95%CI)) of 1.48 (3.98, 25.62) and 31.30 (8.56, 54.04), respectively. For the HEV population aged 0–14 years, GEM shows a significant positive risk effect on the number of cases (βWQS (95%CI): 0.02 (0.01, 0.04)). Conversely, in the HEV population aged 0–14 years, the overall toxic mixture exhibits a significant negative effect on the number of cases (βWQS (95%CI): -3.14 (-6.02, -0.25)).

Table 3
Assessment of the effects of single and mixed toxic substances on the risk of HEV infection in different age groups using Weighted Quantile Sum (WQS) regression.

Series	Mixtures	Single toxics regression survey-weighted		Multiple toxics regression survey-weighted	
		β_{WQS} (95%CI)	p-Value	β_{WQS} (95%CI)	p-Value
Total	PBM	80.60 (20.04, 141.16)	**0.01***	-104.00 (-452.88, 244.88)	0.56
Aged 0–14 years	GEM	0.02 (0.01, 0.04)	**0.000012*****	-3.14 (-6.02, -0.25)	**0.035***
Aged 15–39 years	PBM	1.48 (3.98, 25.62)	**0.0086****	-4.17(-107.95, 24.55)	0.22
Aged 40–59 years	PBM	33.70 (5.67, 61.73)	**0.02***	-41.70 (-201.24, 117.84)	0.61
Aged 60 + years	GEM	-0.66 (-1.13, -0.18)	**0.0084****	-14.70 (-147.59, 118.19)	0.83
	PBM	31.30 (8.56, 54.04)	**0.0078****		
*** P < 0.001, ** P < 0.01, * P < 0.05					
Bold font indicates statistical significance at the 0.05 level.					
Abbreviations: βWQS: Weighted Quantile Sum regression coefficient; 95% CI: 95% confidence interval; GEM: Gaseous elemental mercury; PBM: Particulate-bound mercury.					

The weight of each toxic factor can be determined through subgroup analysis of the WQS mixture. As shown in **Figures S5 A/C/E/G/I**, the sensitive toxic factors that primarily influence the number of HAV cases in the overall population and different subgroups are GOM and PAH_BbF. In contrast, the response relationship of the toxic mixture effect on the number of HAV cases is not significant, as seen in **Figures S5 B/D/F/H/J**. Similarly, **Figures S6 A/C/E/G/I** indicate that the sensitive toxic factors that primarily influence the number of HEV cases in the overall population and different subgroups are GOM and PBM. However, the response relationship of the toxic mixture effect on the number of HEV cases is also not significant, as shown in **Figures S6 B/D/F/H/J**.

3.5 Sensitivity and features of factors

From **Table S3**, the seasonal distribution of each dominant period is as follows: the HAV-dominant period primarily occurs in the summer (28.28%), while the HEV-dominant period mainly occurs in the spring (49.95%). The dual-dominant period is also primarily observed in the summer (28.37%). The temporal distribution of each dominant period is as follows: from 2005 to 2009, the HAV-dominant period predominates; from 2010 to 2016 and at the 2020 time point, the dual-dominant period is

prevalent; and from 2017 to 2019, the HEV-dominant period is predominant. Overall, from 2005 to 2020, the HAV-dominant period is the most significant, comprising 45.36%.

Figure 3 presents the results of the random forest model analysis conducted across different periods (HAV-dominant, HEV-dominant, and Dual-dominant). This analysis systematically evaluates the impact of varying levels of toxicant emissions on the incidence of viral hepatitis from three dimensions: variable importance, interactions, and marginal prediction effects. In the HAV-dominant period (Fig. 3A), the model has an R^2 value of 0.32, with core sensitive factors identified as GOM, PCB, and HCB, which are key influencing factors during this period. These factors demonstrated the highest increase in mean squared error (MSE) and all reached significant levels. In the HEV-dominant period (Fig. 3D), the model's R^2 value is 0.24, with core sensitive factors being GEM, PCDD_F, and HCB. Among these, PCDD_F exhibits the strongest significance ($P < 0.01$). In the Dual-dominant period (Fig. 3G), the model achieves an R^2 value of 0.44, with core sensitive factors identified as GEM, GOM, HCB, PAH_BbF, and PCDD_F. Notably, GEM, GOM, and HCB show the highest increases in MSE, making them the core driving factors during this dual-dominant phase. Overall, following the transition in the hepatitis epidemic type, there are significant differences in environmental toxicant driving factors, and the mixed epidemic phase characterized by dual dominance exhibits complexities related to exposure from environmental toxicant emissions.

3.6 Interaction effect

Figures 3B**/E/H** illustrate the interaction (Mean minimal depth) analysis, where the red horizontal line represents the threshold; interactions below this line are considered strong interactions. In the HAV-dominant period, the interaction between PCB and GOM is the strongest (Fig. 3B), which fully corresponds with the variable importance results from Fig. 3A, confirming the core status of both factors. In the HEV-dominant period, the strongest interaction is observed between GEM and GOM (Fig. 3E). In the Dual-dominant period, the strongest interaction is also between GEM and GOM (Fig. 3H), followed by the interaction between GEM and HCB. The interaction network is more complex during this phase, aligning with the multifactor importance results from Fig. 3G and reflecting the characteristics of combined exposure to multiple toxicants during the mixed epidemic phase.

Figures 3C**/F/I** analyze the marginal prediction effects using Partial Dependence Plots. In the HAV-dominant period, exposure levels to PCB and GOM are significantly positively correlated with the risk of disease (Fig. 3C). In the HEV-dominant period, exposure levels to mercury pollutants (GEM/GOM) are the core driving factors for disease incidence, exhibiting a clear dose-response relationship (Fig. 3F). In the Dual-dominant period, the concentrations of GEM and GOM show a very strong positive correlation with disease risk, and their synergistic effect is the strongest during this phase, making them the core drivers of disease incidence.

4. Discussion

4.1 Prediction Results of Viral Hepatitis

The impact of model time dimension selection on predictive performance exhibits clear outcome specificity, which is one of the core findings of this study. For predicting the incidence of HAV and HEV, the monthly SARIMA model demonstrates significantly superior predictive accuracy compared to the quarterly model. The primary reason for this difference may be that the incidence rates of HAV and HEV are more significantly influenced by short-term fluctuating factors such as seasonal variations, patterns of human activity, and environmental factors (e.g., dietary hygiene, climate)[29]. Monthly data can more sensitively capture these short-term fluctuation characteristics, whereas quarterly data tends to smooth out some of the short-term changes, resulting in reduced prediction accuracy. In contrast, for predicting the mortality rates of HAV and HEV, the quarterly SARIMA model shows a greater advantage, with optimal models corresponding to (1,0,1) (0,1,0) [12] and (1,1,2) (0,1,1) [12]. This may be attributed to the fact that changes in mortality rates are generally more stable than those in incidence rates, being less influenced by short-term random factors and more associated with long-term stable factors such as disease severity, quality of medical care, and the baseline health status of the population[30, 31]. Consequently, quarterly data can better reflect this long-term trend, reducing the interference of short-term random fluctuations on predictive results, thereby enhancing prediction accuracy.

In terms of multifactorial prediction, the optimization results for the number and layers of hidden neurons further validate the core principle that “model structure must align with prediction objectives.” The results of this study indicate that, for both HAV and HEV, a high hidden layer structure (3 layers) demonstrates superior predictive performance in incidence rate predictions, while a low hidden layer structure (1–2 layers) is more suitable for mortality rate predictions. The essence of this pattern is determined by the complexity of the prediction objectives: the incidence rates of hepatitis are influenced by a combination of multiple factors, resulting in more complex variation patterns that require more hidden layers and neurons to capture the nonlinear relationships among these factors. This finding is similar to a study conducted in Nanjing, China, which revealed that the incidence of hand, foot, and mouth disease is affected by various factors such as climate and population movement. By constructing a BP neural network model, the study found that increasing the number of hidden neurons improved the model's fit for incidence rates[32]. In contrast, the changes in mortality rates are relatively stable and less affected by short-term random factors, being more associated with stable factors such as disease severity, thus not requiring a complex hidden layer structure[33, 34].

4.2 Results of Mixed Effects

The results of this study indicate that the association between HAV, HEV incidence and toxic exposure exhibits significant age specificity and toxin specificity, which is closely related to the physiological characteristics, exposure patterns of different age groups, and the epidemiological characteristics of the two types of hepatitis viruses. In the association with HAV incidence, PAH_BbF shows the strongest risk effect, which may be related to the hepatotoxicity of PAHs. Previous studies have confirmed that polycyclic aromatic hydrocarbons can increase the risk of disease following hepatitis virus infection by

interfering with liver metabolism and inducing oxidative stress damage to liver cells[35, 36]. GEM and GOM, as gaseous forms of mercury, can accumulate in the liver after inhalation through the respiratory tract, impairing liver cell function[37, 38] and creating a synergistic pathogenic effect with HAV infection. Among the HAV population aged over 60, only PBM exhibited a significant positive risk effect (βWQS (95%CI): 20.95 (0.63, 41.27)), which is closely related to the physiological characteristics of the elderly. The physiological decline in liver metabolic function in older adults enhances their ability to absorb and accumulate particulate-bound mercury (PBM), and with a decreased immune response, liver cell damage induced by PBM is more likely to trigger HAV incidence[37–39]. In the association with HEV incidence, the age specificity of toxic exposure is even more pronounced: except for the population aged 0–14, PBM has a significant positive impact on HEV incidence across all age groups, with higher significance levels observed in the 15–39 and over 60 age groups ($P < 0.01$). The 15–39 age group has a wider range of activities and more exposure opportunities, while individuals over 60 have reduced immunity. PBM can enter the body through the respiratory tract and skin contact, accumulating in the liver, exacerbating liver cell damage, reducing the body's resistance to HEV, and thereby increasing the risk of disease[40–42].

4.3 Results of Sensitive Factors in Different Dominant Periods

There are significant seasonal and temporal distribution differences in the dominant types of HAV and HEV epidemics, which are closely related to the epidemiological characteristics of the two types of hepatitis viruses and the seasonal fluctuations of environmental factors. The results indicate that HAV-dominant and dual-dominant types are primarily distributed in the summer, while HEV-dominant types are concentrated in the spring. This is consistent with previous research findings that both HAV and HEV are primarily transmitted via the gastrointestinal tract. Higher temperatures and increased humidity in the summer facilitate the survival and spread of the virus in the environment[43, 44]. In contrast, the frequent activities in livestock and poultry farming during the spring, coupled with an increased risk of fecal contamination, are more likely to trigger HEV outbreaks[45–47]. Different years exhibit distinct dominant types of hepatitis, reflecting a clear phase transition characteristic in hepatitis epidemic types[48]. This variation is speculated to be closely related to changes in population immunity levels[49, 50], alterations in environmental toxin emission patterns, and the implementation of prevention and control measures[51].

Random forest model analysis revealed significant differences in the core environmental toxic factors during different dominant hepatitis periods. During the HAV-dominant period, GOM, PCB, and HCB were identified as key influencing factors, all reaching significant levels. This aligns with research findings that persistent organic pollutants such as PCB and HCB can interfere with liver metabolism and induce oxidative stress damage to liver cells[52, 53], thereby increasing the risk of HAV incidence. Additionally, GOM, as a gaseous form of mercury, can accumulate in the liver after inhalation through the respiratory tract, impairing liver cell function[54–56] and creating a synergistic pathogenic effect with HAV infection. In the HEV-dominant period, PCDD_F was identified as the most significant core sensitive factor ($P <$

0.01). Previous studies have confirmed that PCDD_F exhibits strong hepatotoxicity, damaging the barrier function of liver cells and reducing the body's resistance to HEV, thereby increasing the risk of disease[57–60]. During the dual-dominant period, the characteristics of combined exposure to multiple toxins were more pronounced, demonstrating a stronger synergistic driving effect on hepatitis incidence. This also explains why the model had the highest explanatory power during this period. Interaction analysis further revealed the synergistic exposure patterns of toxins during different dominant periods, confirming the core driving role of mercury pollutants in hepatitis epidemics.

4.4 Application of Different Feature Selection Methods

This study employed a combined strategy of Spearman correlation analysis and LASSO regression for environmental toxic factor selection, with each method complementing the other to ensure the reliability of subsequent modeling and driver factor identification. Before building the BP neural network model, Spearman correlation analysis, which does not require the assumption of normal distribution for variables, was used to quickly screen pollutants associated with HAV and HEV incidence and mortality. This approach streamlined the input dimensions to enhance model training efficiency and mitigate the risks of structural redundancy and overfitting[61]. Considering the prevalent multicollinearity among environmental toxins, LASSO regression applied L1 regularization penalties to compress the coefficients of redundant variables to zero, thereby achieving automatic selection of key factors[62]. The optimal λ value was determined through 5-fold cross-validation, further enhancing the robustness and interpretability of the selection results[63]. Additionally, LASSO performed selection stratified by overall population and age groups, aligning with the heterogeneous characteristics of the population. This provided a reliable variable foundation for subsequent mixed exposure analysis and temporal driver factor identification, ensuring consistency and comparability across multiple model results[64].

4.5 Limitations and Prospects

Nevertheless, this study has several limitations. The prediction models only included chemical pollutants and time-series information without adjusting for confounding factors such as population health status, medical resources, and seasonal patterns, which may lead to bias; in addition, the structure of the BP neural network was not systematically optimized, and long-term predictive stability was not verified. Future studies should incorporate more comprehensive covariates, optimize model parameters, and conduct external validation to improve accuracy and applicability. Moreover, our findings identified age-specific and period-specific heterogeneous effects of mercury species and UPOPs on HAV and HEV, with distinct key toxicants across different epidemic phases, supporting targeted pollution control and hepatitis prevention strategies. Further research combining high-resolution spatial data from EDGAR is warranted to explore regional disparities and underlying mechanisms of multi-pollutant interactions, thus providing more precise evidence for environmental health interventions.

5. Conclusion

This study confirms that SARIMA and BP neural networks can effectively achieve accurate predictions of HAV and HEV incidence and mortality rates, with prediction accuracy influenced by the time scale and model structure. Environmental toxins such as gaseous mercury, particulate mercury, polycyclic aromatic hydrocarbons, and polychlorinated biphenyls are key risk factors affecting the incidence of hepatitis A and E. There are differences in susceptibility to environmental toxins among different age groups; for instance, the elderly HEV population is more sensitive to exposure to particulate mercury, while children in the HEV population are primarily affected by exposure to elemental mercury in its gaseous form. Changes in the dominant type of hepatitis epidemic are associated with alterations in environmental driving factors, with the dual-dominant period displaying complex risk characteristics due to multi-toxin synergistic exposure. Incorporating environmental pollutant indicators into infectious disease prediction models can significantly enhance the accuracy of early warnings for hepatitis incidence and mortality. The results of this study provide important data support for developing targeted hepatitis prevention and control strategies and environmental health management plans based on population and temporal factors.

Declarations

Ethics approval and consent to participate

This study involved no human experimentation. Data used in the present study were obtained from the National Public Health Science Data Center and contained no personal or health information that could be linked to original identifiers. All data were anonymized before utilization in this study.

Consent for publication

All authors have consented to publication of this research.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The viral hepatitis data was publicly supported from the National Public Health Data Centre of China (https://www.phsciencedata.cn/). Mercury and UPOPs emissions information were publicly from the Emissions Database for Global Atmospheric Research (EDGAR) (https://edgar.jrc.ec.europa.eu/).

Availability of data and material

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 the Emissions Database for Global Atmospheric Research (EDGAR) (https://edgar.jrc.ec.europa.eu/).

CRediT authorship contribution statement

Weiming Hou: Conceptualization, Methodology, Formal analysis, Writing-original draft, 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

None.

Clinical trial number

Not applicable.

References

1. Jacobsen KH, Koopman JS: **Declining hepatitis A seroprevalence: a global review and analysis**. *Epidemiol Infect* 2004, **132**(6):1005-1022.
2. Tominaga A, Kanda T, Akiike T, Komoda H, Ito K, Abe A, Aruga A, Kaneda S, Saito M, Kiyohara T *et al*: **Hepatitis A outbreak associated with a revolving sushi bar in Chiba, Japan: Application of molecular epidemiology**. *Hepatol Res* 2012, **42**(8):828-834.
3. Tsukada R, Ono S, Kobayashi H, Wada Y, Nishizawa K, Fujii M, Takeuchi M, Kuroiwa K, Kobayashi Y, Ishii K *et al*: **A Cluster of Hepatitis A Infections Presumed to be Related to Asari Clams and Investigation of the Spread of Viral Contamination from Asari Clams**. *Jpn J Infect Dis* 2019, **72**(1):44-48.
4. Bai H, Shiota T, Yoshizaki S, Saito-Obata M, Malbas FF, Jr., Lupisan SP, Oshitani H, Takeda N, Muramatsu M, Wakita T *et al*: **Detection of Subgenotype IA and IIIA Hepatitis A Viruses in Rivers Flowing through Metro Manila, the Philippines**. *Jpn J Infect Dis* 2019, **72**(1):53-55.
5. Joon A, Rao P, Shenoy SM, Baliga S: **Prevalence of Hepatitis A virus (HAV) and Hepatitis E virus (HEV) in the patients presenting with acute viral hepatitis**. *Indian J Med Microbiol* 2015, **33 Suppl**:102-105.
6. Wang B, Meng XJ: **Structural and molecular biology of hepatitis E virus**. *Comput Struct Biotechnol J* 2021, **19**:1907-1916.
7. Kanda T, Li TC, Takahashi M, Nagashima S, Primadharsini PP, Kunita S, Sasaki-Tanaka R, Inoue J, Tsuchiya A, Nakamoto S *et al*: **Recent advances in hepatitis E virus research and the Japanese clinical practice guidelines for hepatitis E virus infection**. *Hepatol Res* 2024, **54**(8):1-30.
8. Shin EC, Jeong SH: **Natural History, Clinical Manifestations, and Pathogenesis of Hepatitis A**. *Cold Spring Harb Perspect Med* 2018, **8**(9).
9. Kamar N, Dalton HR, Abravanel F, Izopet J: **Hepatitis E virus infection**. *Clin Microbiol Rev* 2014, **27**(1):116-138.

10. Song X, Xiao J, Deng J, Kang Q, Zhang Y, Xu J: **Time series analysis of influenza incidence in Chinese provinces from 2004 to 2011**. *Medicine (Baltimore)* 2016, **95**(26):e3929.
11. Kane MJ, Price N, Scotch M, Rabinowitz P: **Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks**. *BMC Bioinformatics* 2014, **15**(1):276.
12. Xu Q, Gel YR, Ramirez Ramirez LL, Nezafati K, Zhang Q, Tsui KL: **Forecasting influenza in Hong Kong with Google search queries and statistical model fusion**. *PLoS One* 2017, **12**(5):e0176690.
13. Wang KW, Deng C, Li JP, Zhang YY, Li XY, Wu MC: **Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network**. *Epidemiol Infect* 2017, **145**(6):1118-1129.
14. 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* 2023, **17**(7):e0010806.
15. Liang Y, Lin X, Zhang M, Wang S, Dong Y, Xie Z, He X, Ju W, Chen M: **Synergistic impact of high body-mass index on HBV/HCV-related liver cancer burden globally: Analysis and projections 1990-2050**. *Public Health* 2025, **247**:105893.
16. Hou W, Song Z: **Exploring the risk and predictive study of outdoor air pollutants on the incidence and mortality of HIV/AIDS**. *Ecotoxicol Environ Saf* 2024, **287**:117292.
17. Li Y, Zhou T, Liu Z, Zhu X, Wu Q, Meng C, Deng Q: **Air pollution and prostate cancer: Unraveling the connection through network toxicology and machine learning**. *Ecotoxicol Environ Saf* 2025, **292**:117966.
18. Navarro-Sempere A, García M, Rodrigues AS, Garcia PV, Camarinho R, Segovia Y: **Occurrence of Volcanogenic Inorganic Mercury in Wild Mice Spinal Cord: Potential Health Implications**. *Biol Trace Elem Res* 2022, **200**(6):2838-2847.
19. Cariccio VL, Samà A, Bramanti P, Mazzon E: **Mercury Involvement in Neuronal Damage and in Neurodegenerative Diseases**. *Biol Trace Elem Res* 2019, **187**(2):341-356.
20. Goodrich JA, Wang H, Jia Q, Stratakis N, Zhao Y, Maitre L, Bustamante M, Vafeiadi M, Aung M, Andrušaitytė S *et al*: **Integrating Multi-Omics with environmental data for precision health: A novel analytic framework and case study on prenatal mercury induced childhood fatty liver disease**. *Environ Int* 2024, **190**:108930.
21. Hakkarainen K, Rantakokko P, Koponen J, Ruokojärvi P, Korkalainen M, Salomaa V, Jula A, Männistö S, Perola M, Lundqvist A *et al*: **Persistent organic pollutants associate with liver disease in a Finnish general population sample**. *Liver Int* 2023, **43**(10):2177-2185.
22. Joshi-Barve S, Kirpich I, Cave MC, Marsano LS, McClain CJ: **Alcoholic, Nonalcoholic, and Toxicant-Associated Steatohepatitis: Mechanistic Similarities and Differences**. *Cell Mol Gastroenterol Hepatol* 2015, **1**(4):356-367.
23. Letcher RJ, Bustnes JO, Dietz R, Jenssen BM, Jørgensen EH, Sonne C, Verreault J, Vijayan MM, Gabrielsen GW: **Exposure and effects assessment of persistent organohalogen contaminants in arctic wildlife and fish**. *Sci Total Environ* 2010, **408**(15):2995-3043.

24. Li G, Li Y, Han G, Jiang C, Geng M, Guo N, Wu W, Liu S, Xing Z, Han X *et al*: **Forecasting and analyzing influenza activity in Hebei Province, China, using a CNN-LSTM hybrid model**. *BMC Public Health* 2024, **24**(1):2171.
25. Yang Y, Hu R, Wang W, Zhang T: **Construction and optimization of non-parametric analysis model for meter coefficients via back propagation neural network**. *Sci Rep* 2024, **14**(1):11452.
26. He C, Zhang Y, Tang M, Ai X, Tang M, Tang C, Li L, Huang W, You X, Zhou D *et al*: **Analysis of barriers associated with emergency medical service activation in patients with acute stroke and acute myocardial infarction from Zhongjiang County of Sichuan Province in China**. *BMC Emerg Med* 2024, **24**(1):113.
27. Jin Z, Sun W, Huang J, Wang G: **Heavy metals and their relationships with lung function and airway inflammation: insights from a population-based study**. *J Transl Med* 2025, **23**(1):642.
28. Zheng J, Tao L, Dini-Andreote F, Luan L, Kong P, Xue J, Zhu G, Xu Q, Jiang Y: **Dynamic Responses of Ammonia-Oxidizing Archaea and Bacteria Populations to Organic Material Amendments Affect Soil Nitrification and Nitrogen Use Efficiency**. *Front Microbiol* 2022, **13**:911799.
29. Huang X, Zheng X, Chen L, Zhuang C, Yang C, Zang X, Wang Y, Jiang H, Huang X, Yan Q *et al*: **Active surveillance of hepatitis E: a 10-year epidemiological analysis in a city in eastern China**. *Emerg Microbes Infect* 2024, **13**(1):2373315.
30. Guo F, Ma X, Shan Q, Zhang Y, Gao S, Chen J, Chen Y, Liu D, Li T, Zhang F *et al*: **Epidemiological insights into severe hepatitis in pediatric inpatients**. *iScience* 2024, **27**(12):111420.
31. Lourens S, Sunjaya DB, Singal A, Liangpunsakul S, Puri P, Sanyal A, Ren X, Gores GJ, Radaeva S, Chalasani N *et al*: **Acute Alcoholic Hepatitis: Natural History and Predictors of Mortality Using a Multicenter Prospective Study**. *Mayo Clin Proc Innov Qual Outcomes* 2017, **1**(1):37-48.
32. Liu W, Bao C, Zhou Y, Ji H, Wu Y, Shi Y, Shen W, Bao J, Li J, Hu J *et al*: **Forecasting incidence of hand, foot and mouth disease using BP neural networks in Jiangsu province, China**. *BMC Infect Dis* 2019, **19**(1):828.
33. Shanbehzadeh M, Nopour R, Kazemi-Arpanahi H: **Design of an artificial neural network to predict mortality among COVID-19 patients**. *Inform Med Unlocked* 2022, **31**:100983.
34. Lehman LW, Ghassemi M, Snoek J, Nemati S: **Patient Prognosis from Vital Sign Time Series: Combining Convolutional Neural Networks with a Dynamical Systems Approach**. *Comput Cardiol (2010)* 2015, **42**:1069-1072.
35. Kim M, Jee SC, Sung JS: **Hepatoprotective Effects of Flavonoids against Benzo[a]Pyrene-Induced Oxidative Liver Damage along Its Metabolic Pathways**. *Antioxidants (Basel)* 2024, **13**(2).
36. Jin Y, Miao W, Lin X, Pan X, Ye Y, Xu M, Fu Z: **Acute exposure to 3-methylcholanthrene induces hepatic oxidative stress via activation of the Nrf2/ARE signaling pathway in mice**. *Environ Toxicol* 2014, **29**(12):1399-1408.
37. Khayat A, Dencker L: **Whole body and liver distribution of inhaled mercury vapor in the mouse: influence of ethanol and aminotriazole pretreatment**. *J Appl Toxicol* 1983, **3**(2):66-74.

38. Yahyazedeh A, Altunkaynak BZ, Akgül N, Akgül HM: **A histopathological and stereological study of liver damage in female rats caused by mercury vapor**. *Biotech Histochem* 2017, **92**(5):338-346.
39. He QJ, Li YF, Zhao LT, Lin CT, Yu CY, Wang D: **Recent advances in age-related metabolic dysfunction-associated steatotic liver disease**. *World J Gastroenterol* 2024, **30**(7):652-662.
40. Li P, Liu J, Li Y, Su J, Ma Z, Bramer WM, Cao W, de Man RA, Peppelenbosch MP, Pan Q: **The global epidemiology of hepatitis E virus infection: A systematic review and meta-analysis**. *Liver Int* 2020, **40**(7):1516-1528.
41. Wu JY, Lau EHY, Lu ML, Guo C, Guo ZM, Yuan J, Lu JH: **An occupational risk of hepatitis E virus infection in the workers along the meat supply chains in Guangzhou, China**. *One Health* 2022, **14**:100376.
42. Campbell R, Russell C, Izopet J, Nassim K, Sonderup M, Pischke S, Wallace SJ, Madden RG: **Environmental considerations in hepatitis E virus transmission: Is there a missing link?** *World J Hepatol* 2026, **18**(2):112467.
43. Villar LM, De Paula VS, Gaspar AM: **Seasonal variation of hepatitis A virus infection in the city of Rio de Janeiro, Brazil**. *Rev Inst Med Trop Sao Paulo* 2002, **44**(5):289-292.
44. Parashar D, Khalkar P, Arankalle VA: **Survival of hepatitis A and E viruses in soil samples**. *Clin Microbiol Infect* 2011, **17**(11):E1-4.
45. Turlewicz-Podbielska H, Augustyniak A, Wojciechowski J, Pomorska-Mól M: **Hepatitis E Virus in Livestock-Update on Its Epidemiology and Risk of Infection to Humans**. *Animals (Basel)* 2023, **13**(20).
46. Lu YH, Qian HZ, Hu AQ, Qin X, Jiang QW, Zheng YJ: **Seasonal pattern of hepatitis E virus prevalence in swine in two different geographical areas of China**. *Epidemiol Infect* 2013, **141**(11):2403-2409.
47. Hofmeister MG, Foster MA, Teshale EH: **Epidemiology and Transmission of Hepatitis A Virus and Hepatitis E Virus Infections in the United States**. *Cold Spring Harb Perspect Med* 2019, **9**(4).
48. Ren X, Wu P, Wang L, Geng M, Zeng L, Zhang J, Xia N, Lai S, Dalton HR, Cowling BJ *et al*: **Changing Epidemiology of Hepatitis A and Hepatitis E Viruses in China, 1990-2014**. *Emerg Infect Dis* 2017, **23**(2):276-279.
49. Migueres M, Lhomme S, Izopet J: **Hepatitis A: Epidemiology, High-Risk Groups, Prevention and Research on Antiviral Treatment**. *Viruses* 2021, **13**(10).
50. Jacobsen KH: **Globalization and the Changing Epidemiology of Hepatitis A Virus**. *Cold Spring Harb Perspect Med* 2018, **8**(10).
51. Mathur P, Arora NK: **Epidemiological transition of hepatitis A in India: issues for vaccination in developing countries**. *Indian J Med Res* 2008, **128**(6):699-704.
52. Pan S, Li Z, Rubbo B, Quon-Chow V, Chen JC, Baumert BO, Garcia E, Aung MT, Conti DV, Chatzi L: **Applications of mixture methods in epidemiological studies investigating the health impact of persistent organic pollutants exposures: a scoping review**. *J Expo Sci Environ Epidemiol* 2025, **35**(4):522-534.

53. Tian Y, Rimal B, Gui W, Koo I, Yokoyama S, Perdew GH, Patterson AD: **Early Life Short-Term Exposure to Polychlorinated Biphenyl 126 in Mice Leads to Metabolic Dysfunction and Microbiota Changes in Adulthood**. *Int J Mol Sci* 2022, **23**(15).
54. Yeh M-J, Yuan C-S, Hung K-N, Ie I-R, Lee C-E, Chiang K-C, Soong K-Y: **Temporal variation and potential origins of atmospheric speciated mercury at a remote island in South China Sea based on two-year field measurement data**. *Scientific Reports* 2021, **11**(1):5678.
55. Banerjee P, Gaddam N, Chandler V, Chakraborty S: **Oxidative Stress-Induced Liver Damage and Remodeling of the Liver Vasculature**. *Am J Pathol* 2023, **193**(10):1400-1414.
56. Choi J, Bae S, Lim H, Lim JA, Lee YH, Ha M, Kwon HJ: **Mercury Exposure in Association With Decrease of Liver Function in Adults: A Longitudinal Study**. *J Prev Med Public Health* 2017, **50**(6):377-385.
57. Li C, Liu Y, Dong Z, Xu M, Gao M, Cong M, Liu S: **TCDD promotes liver fibrosis through disordering systemic and hepatic iron homeostasis**. *J Hazard Mater* 2020, **395**:122588.
58. Cong Y, Hong Y, Wang D, Cheng P, Wang Z, Xing C, Sun W, Xu G: **2,3,7,8-Tetrachlorodibenzo-p-dioxin induces liver lipid metabolism disorder via the ROS/AMPK/CD36 signaling pathway**. *Toxicol Sci* 2023, **191**(2):276-284.
59. Pham PQ, Nguyen VB, Pham TT, Duong NX, Nguyen HT, Ha QV, Nguyen TD, Hoang TM, Dinh DT, Tran QTN *et al*: **Histopathological Alterations in the Livers of Chronic Hepatitis Patients Exposed to Agent Orange/Dioxin in Vietnam**. *Toxics* 2022, **10**(6).
60. Yan J, Tung HC, Li S, Niu Y, Garbacz WG, Lu P, Bi Y, Li Y, He J, Xu M *et al*: **Aryl Hydrocarbon Receptor Signaling Prevents Activation of Hepatic Stellate Cells and Liver Fibrogenesis in Mice**. *Gastroenterology* 2019, **157**(3):793-806.e714.
61. Pei S, Yang L, Gao H, Liu Y, Dai E, Feng F, Lu J: **Pollutants-mediated viral hepatitis in different types: assessment of different algorithms and time series models**. *Sci Rep* 2024, **14**(1):21141.
62. Xi LJ, Guo ZY, Yang XK, Ping ZG: **[Application of LASSO and its extended method in variable selection of regression analysis]**. *Zhonghua Yu Fang Yi Xue Za Zhi* 2023, **57**(1):107-111.
63. Zhu G, Wen Y, Cao K, He S, Wang T: **A review of common statistical methods for dealing with multiple pollutant mixtures and multiple exposures**. *Front Public Health* 2024, **12**:1377685.
64. Handorf E, Yin Y, Slifker M, Lynch S: **Variable selection in social-environmental data: sparse regression and tree ensemble machine learning approaches**. *BMC Med Res Methodol* 2020, **20**(1):302.

Figures

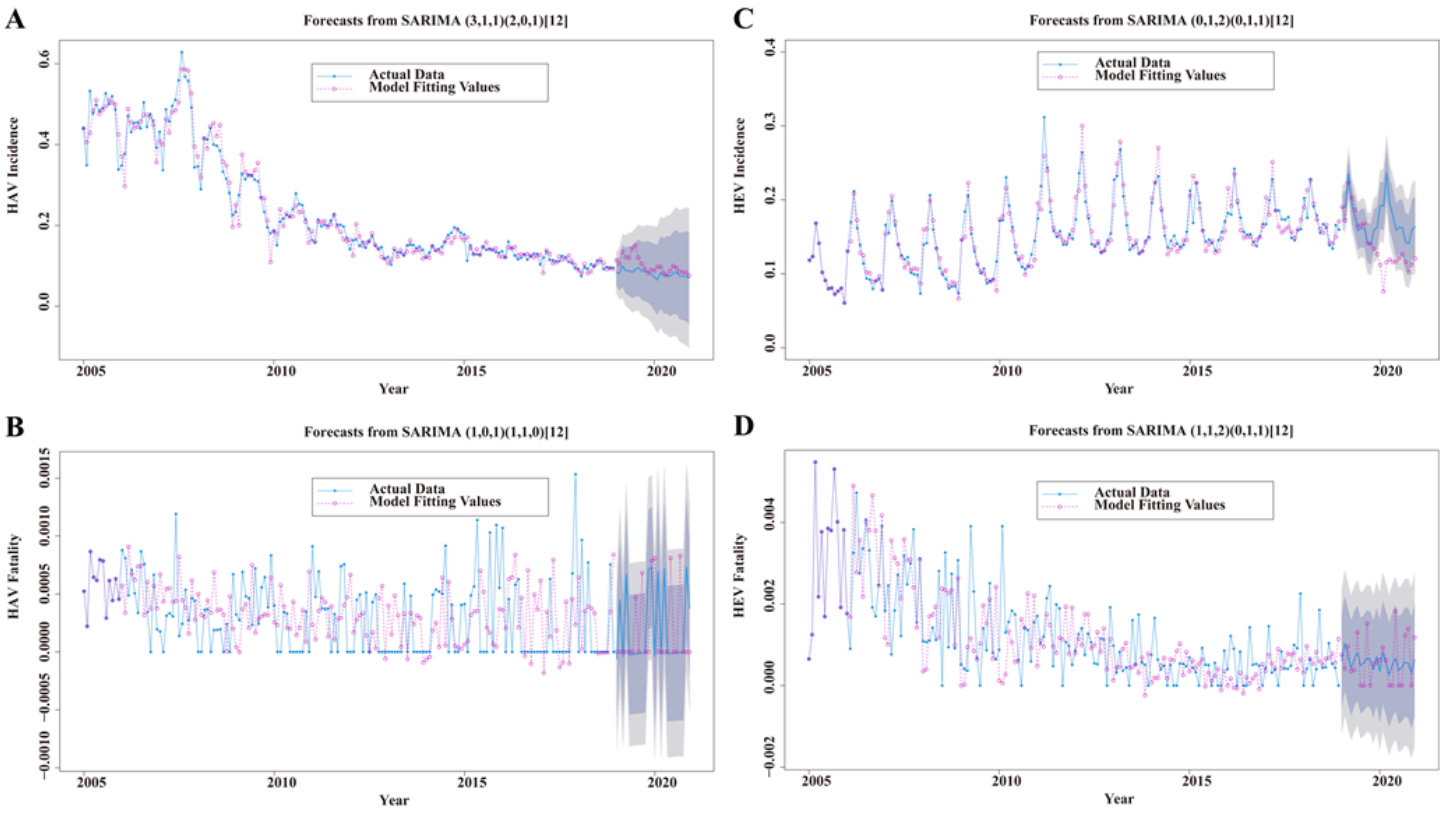


Figure 1

Time series fitting and forecasting of HAV and HEV incidence and mortality rates based on the SARIMA monthly model. A: HAV incidence; B: HAV mortality rate; C: HEV incidence; D: HEV mortality rate; the blue solid line represents actual values, the pink dashed line represents model fitted values, and the gray area represents the 95% prediction confidence interval.

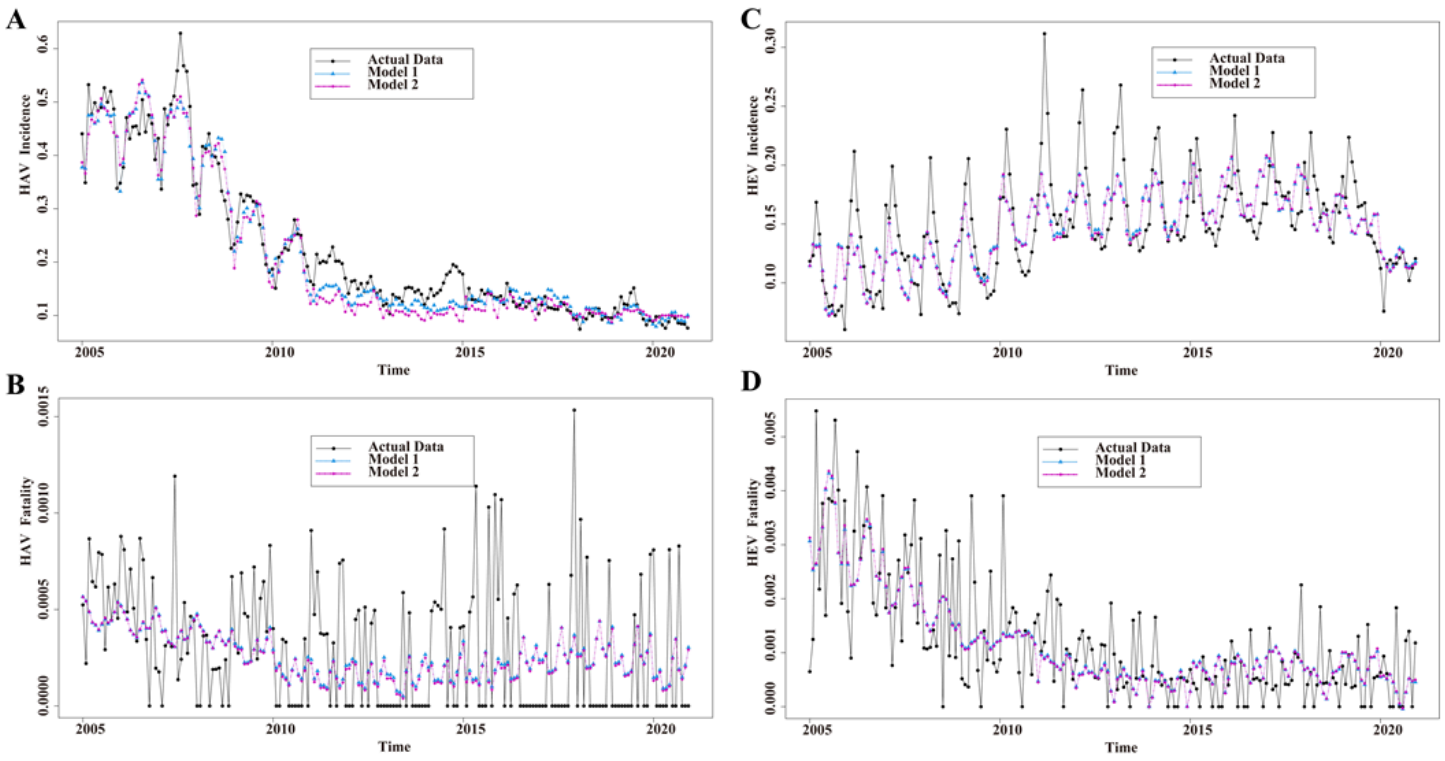


Figure 2

Prediction fitting of HAV and HEV incidence and mortality using BP neural network models based on different activation functions. A: HAV incidence; B: HAV mortality rate; C: HEV incidence; D: HEV mortality rate; the black solid line represents actual values, the blue dashed line represents the fitted values of Model 1 (hidden layer Sigmoid + output layer purelin), and the pink solid line represents the fitted values of Model 2 (hidden layer Sigmoid + output layer tansig).

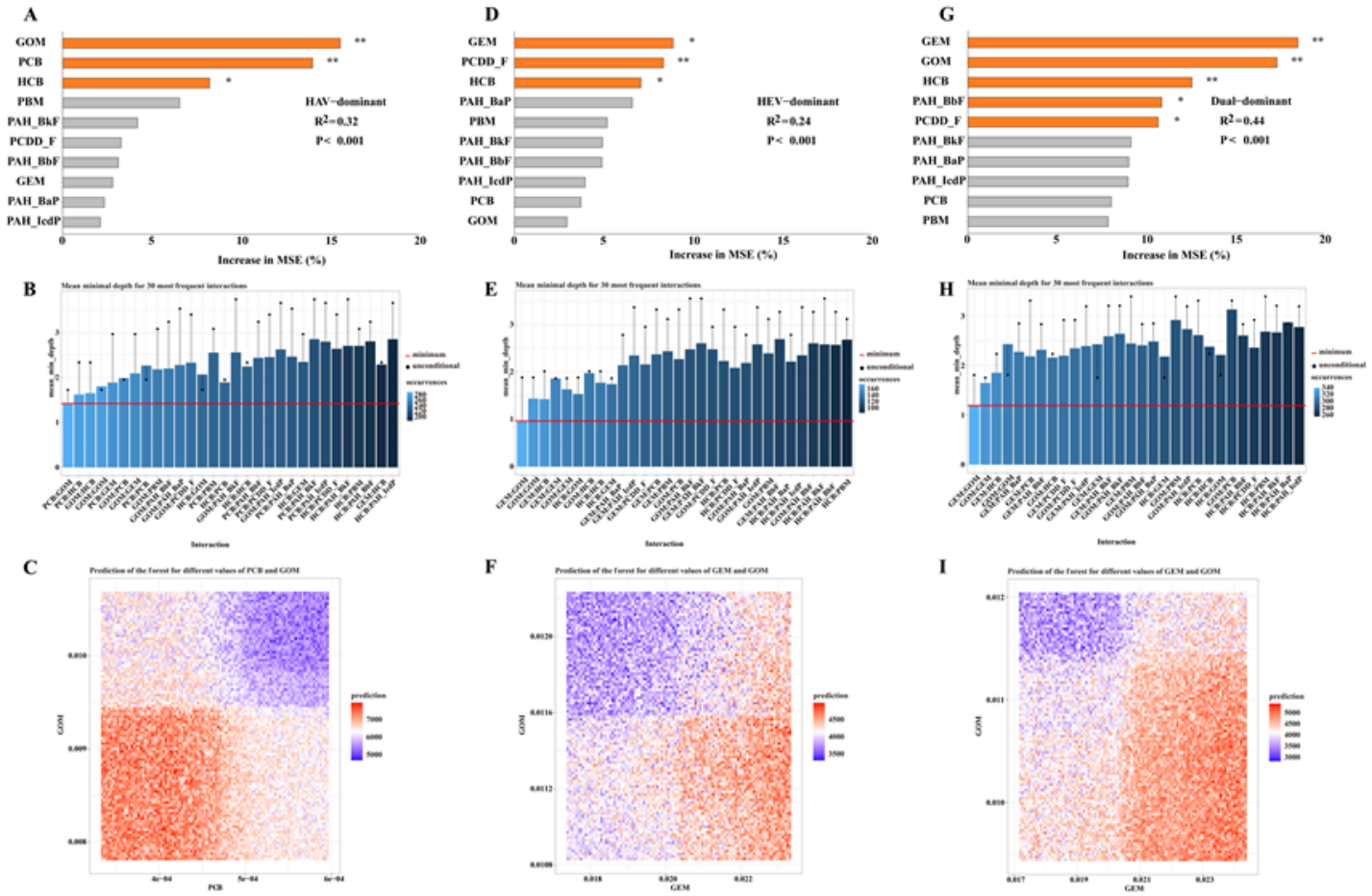


Figure 3

Analysis of the importance, interactions, and marginal effects of environmental toxic factors during different dominant periods of hepatitis epidemics. A-C: HAV dominant period (HAV-dominant); D-F: HEV dominant period (HEV-dominant); G-I: dual dominant period (Dual-dominant); A/D/G: percentage increase in mean squared error (Increase in MSE) showing variable importance; B/E/H: mean minimal depth analyzing interaction intensity; C/F/I: heatmap displaying the dose-response predictive effects of core toxicant combinations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementfiles.doc