Spatiotemporal Analysis for the Effect of Hydraulic Fracturing and Stroke in the United States

Chuanbo Hu
West Virginia University

Bin Liu
West Virginia University

Shuo Wang
Washington University in St. Louis

Zhenduo Zhu
State University of New York

Xin Li (✉ xin.li@mail.wvu.edu)
West Virginia University

Spatiotemporal Analysis for the Effect of Hydraulic Fracturing and Stroke in the United States

Chuanbo Hu[1], Bin Liu[2], Shuo Wang[3], Zhenduo Zhu[4], and Xin Li[1,*]

[1]Lane Department of CSEE, West Virginia University, PO Box 6109, Morgantown WV 26506-6109
[2]Department of Management Information System, West Virginia University, PO Box 6025, Morgantown WV 26506-6025
[3]Department of Radiology, Washington University at St. Louis, Campus Box 8225, 4525 Scott Ave, St. Louis, MO 63110
[4]Department of Civil, Structural and Environmental Engineering, State University of New York at Buffalo, 207C Jarvis Hall, Buffalo, NY 14260
[*]Corresponding author. Email: xin.li@mail.wvu.edu

ABSTRACT

Hydraulic fracturing or fracking has led to a rapid growth of oil and gas production in the United States, but the impact of fracking on public health is an important but under-researched topic. We design a methodology to study spatio-temporal correlations between the risk of fracking and stroke mortality. An annualized loss expectancy (ALE) model is applied to quantify the risk of fracking. The Geographically and Temporally Weighted Regression (GTWR) model is used to analyze spatio-temporal correlations of stroke mortality, fracking ALE, and nine other socioeconomic and health-related factors. The analysis shows that ALE from fracking is moderately correlated with stroke mortality at ages over 65 in most states of fracking, in addition to cardiovascular disease and drug overdose being positively correlated with stroke mortality. Furthermore, the correlations between fracking ALE and stroke mortality in men appear to be higher than in women near the Marcellus Shale, including Ohio, Pennsylvania, West Virginia, and Virginia, while stroke mortality among women is concentrated in the Great Plains, including Montana, Wyoming, New Mexico, and Oklahoma. Lastly, within two kilometers of the fracking mining activity, it was found that the level of benzene in the air was significantly correlated with the fracking activity in Colorado.

Introduction

Hydraulic fracturing[1], also known as fracking, is a geochemical engineering process by which large volumes of water combined with chemical and sand proppants are injected into tight formations with high pressure to fracture and facilitate the recovery of unconventional reserves of oil and gas[2]. With the development of fracking technology, the United States (US) is the world's largest producer of both natural gas and crude oil. According to the US Energy Information Administration (EIA), there were approximately 23,000 fracking wells in the US in 2000. In 2015, the number of fracking wells increased rapidly to approximately 300,000, representing 67% of US natural gas production and 51% of US[3]. Despite the economic benefits of fracking, this expansion has brought the industrial activities of oil and gas development closer to backyards and communities, increasing the risk of human exposure to new contaminants and threats[4,5].

The health implications and effects of fracking have not been adequately studied[6–11]. There is a growing body of research studies on the negative impact of fracking on air and water quality[12–16], as well as public health[17–21]. Although health discussions have focused on drinking water contamination, particularly in the eastern US, there is growing interest in studying a variety of health threats arising from air pollution[6,22]. For example, research has attempted to link fracking pollution with unhealthy levels of smog and toxic air pollutants[23]. Exposure to this pollution can cause eye, nose, throat, respiratory disease, birth defects, cancer, or premature death[24]. However, little is known about whether fracking can cause life-threatening conditions such as stroke[7].

Stroke as a neurological disease is the leading cause of long-term adult disability and the fifth leading cause of death in the US[25], with approximately 795,000 stroke events annually. Stroke Belt refers to a consistent pattern of striking geographic variation in stroke mortality rates within the US[26]. It covers 11 states in the southeast US with an unusually high incidence of stroke and other forms of cardiovascular disease. Factors that explain the prevalence of excessive stroke in the stroke belt include differences in socioeconomic status (e.g., employment rate and marital rate), risk factors (e.g., smoking and unhealthy diet), and prevalence of common chronic diseases (e.g., diabetes and heart disease)[27,28]. A recent study has shown that the highest contributors to the Stroke Belt include a higher burden of risk factors, higher levels of inflammation and infection, and

lower socioeconomic status, while environmental exposures and lifestyle choices are considered lesser contributors[29].

Does fracking induce a higher risk of stroke? Although a potential connection between fracking and stroke has been mentioned in the literature (e.g.,[30–32]), there is no systematic study to address this question. The closest research to this work is the study on the impact of fracking on water pollution[33] and air pollution[34], but its research data are limited to the local area. Furthermore, the spatial extent of the public health impact of hydraulic fracturing is a question that existing research attempts to answer. For example, the distance to the nearest fracking well as an important indicator has been used to analyze the spatial correlation between fracking and infant health[35]. However, few studies have analyzed the spatiotemporal impacts of hydraulic fracturing on public health. To our knowledge, this article is the first attempt to address this question from a geographic information system (GIS) perspective using the extension of the geographically weighted regression (GWR) method[36]: geographically and temporally weighted regression (GTWR)[37]. We performed a detailed regression analysis of the stroke and fracking data using GTWR. We hope that this work can shed light on the relations between fracking and stroke risk and stimulate more quantitative studies on the health risk of fracking, which can better inform decision makers about energy and public health policy.

Results

Model Comparison

To solve the multicollinearity problem between multiple variables, the variables with a VIF greater than 10 (i.e., hypertension and obesity) were removed. The results of VIF values of significant explanatory variables are given in the Tab. 6 for stroke mortality over 65 years and Tab. 7 stroke mortality $45 - 64$ in Appendix A. Additionally, Moran's I statistics were calculated to determine whether the explanatory variables in Tab. 6 and Tab. 7 are spatially associated. The results of the Moran's I test are given in Tab. 8 in the Appendix A. The 10 selected variables with a p-value below 0.05 were included in the regression model, which represents that all variables are spatially autocorrelated.

Moreover, to increase the significance of the regression variable, two explanatory variables (i.e., heavy drink and education) have been removed using stepwise selection based on AICc. In addition, a comparison with three baseline models was implemented, including Ordinary Least Squares (OLS)[38], Temporally Weighted Regression (TWR)[37] and GWR[36]) to evaluate the performance of the GTWR model[37]. As shown in Tab. 1. GTWR outperforms OLS, TWR, and GWR in model fitting, demonstrating that it better explains the dependent variable stroke mortality. Taking the fracking state model as an example, the values of R^2 increase from 0.757 in the OLS model, 0.768 in the TWR model, and 0.933 in the GWR model, to 0.970 in the GTWR model. The AICc reduce from -401.129 in the OLS, -399.780 in the TWR model and -534.502 in the GWR model, to -564.090 in the GTWR model (the lower, the better). Explanatory power increases significantly, since spatial information and temporal information are considered in the model. In the rest of the paper, we will only analyze the results of the GTWR model.

Table 1. Comparison results of OLS, TWR, GWR, and GTWR models.

	Fracking States			Non-fracking States		
	AICc	R^2	*Adjusted R^2*	*AICc*	R^2	*Adjusted R^2*
OLS[38]	-401.129	0.757	0.745	-293.313	0.601	0.584
TWR[37]	-399.780	0.768	0.757	-318.285	0.691	0.678
GWR[36]	-534.502	0.933	0.929	-481.016	0.897	0.892
GTWR[37]	**-564.090**	**0.970**	**0.968**	**-487.886**	**0.931**	**0.928**

The estimations of regression coefficients for 65+ stroke mortality in fracking states and non-fracking states are given in the Tab. 2 and Tab. 3. The results of both models showed the positive effect of cardiovascular and overdose on 65+ stroke mortality, and the negative effect of marital rate on 65+ stroke mortality. These results share some similarities with other findings of previous work[39,40]. In addition, Tab. 2 shows that most fracking ALE coefficients are positively associated with stroke mortality at 65 years, although the correlation was much smaller than for stroke mortality variables such as cardiovascular disease and overdose.

Spatiotemporal Features of Fracking ALE Coefficients

We analyze the temporal and spatial characteristics of fracking ALE using the average values of the regression coefficients, which help to explore the temporal trends and spatial differences of fracking ALE on stroke mortality at ages 65 or older.

Temporal Features of Fracking ALE Coefficients

The aforementioned improvement of GTWR is extended by incorporating the temporal dimension into the traditional GWR model. From the results of the GTWR model, we can obtain the time series of the yearly fracking ALE coefficients. Fig. 1 presents the fluctuation of the average coefficients of the Fracking ALE variables over a period of 9 years (from 2010 to 2018).

Table 2. Estimation of the GTWR model for 65+ stroke mortality in fracking states

Variables	MIN	LQ	MED	UQ	MAX	AVG
Related-disease risk factors						
Diabetes	-0.661	-0.316	-0.001	0.218	0.632	-0.018
Cardiovascular	-1.506	-0.070	0.269	0.524	1.141	0.232
Overdose	-0.861	-0.030	0.389	0.693	1.148	0.347
Behavior risk factors						
Tobacco use	-0.973	-0.342	-0.157	-0.035	0.351	-0.197
High cholesterol	-1.346	-0.325	-0.149	0.090	0.415	-0.139
PAI	-0.585	-0.161	-0.080	0.020	0.521	-0.056
Socioeconomic risk factors						
Mean income	-1.442	-0.215	0.147	0.367	0.623	0.039
Marital rate	-1.459	-0.984	-0.677	-0.409	0.720	-0.663
Employment rate	-2.106	-0.604	-0.014	0.128	0.627	-0.223
Fracking risk factor						
Fracking ALE	-0.327	0.041	0.116	0.154	0.394	0.094
Intercept	-0.009	0.487	0.669	0.792	3.747	0.715

Table 3. Estimation of the GTWR model for 65+ stroke mortality in non-fracking states

Variables	MIN	LQ	MED	UQ	MAX	AVG
Related-disease risk factors						
Diabetes	-0.603	-0.085	0.227	0.314	0.467	0.109
Cardiovascular	-0.354	0.272	0.357	0.427	0.896	0.305
Overdose	-1.341	-0.189	0.166	0.286	0.805	0.037
Behavior risk factors						
Tobacco use	-0.577	-0.036	0.102	0.222	0.71	0.08
High cholesterol	-0.841	-0.539	-0.296	-0.165	0.175	-0.333
PAI	-0.493	-0.101	0.043	0.233	0.451	0.063
Socioeconomic risk factors						
Mean income	-2.403	-0.684	-0.412	-0.222	0.476	-0.479
Marital rate	-1.448	-0.651	-0.078	0.265	0.354	-0.221
Employment rate	-1.657	-1.07	-0.615	-0.144	1.122	-0.507

Negative coefficients indicate the reverse correlation between the dependent and explanatory variables, and vice versa. 19 states with active fracking activity were discussed and 5 states with very few fracking activities were removed[41]. As Fig. 1 shows, the positive correlation between fracking ALE and stroke mortality (65+) in California, Utah, Alabama, Louisiana, and Oklahoma decreased significantly year by year. In contrast, the positive correlation between fracking ALE and stroke mortality (65+) in North Dakota increased significantly. The coefficient of Fracking ALE in West Virginia, Ohio, Pennsylvania, Michigan, and Virginia slowly decreases. The ALE coefficient for fracking in Colorado and New Mexico first increases and then decreases.

Spatial Features of Fracking ALE Coefficients
An important feature of the GTWR model is that the estimated coefficients are mappable for visual analysis. The spatial distributions of the effects of explanatory variables on ALE fracking are visualized in Fig. 2. This study sets zero as a threshold to distinguish positive and negative effects. Darker states on the map have stronger positive correlations between fracking ALE and stroke mortality above 65. Fig. 2 shows that the positive correlation between ALE of fracking and stroke mortality (65+) in North Dakota, Ohio, Montana, Kansas and Arkansas is stronger than in other states of fracking. Fracking has been active in North Dakota and Ohio[41]. Although fracking is generally active in Montana, Kansas, and Arkansas, fracking in their respective neighbors (e.g., North Dakota, Colorado, Texas) is always active. On the contrary, the positive correlation in Virginia, Mississippi, and Oklahoma is weaker than in other states of fracking. Fracking is not active in the states of Virginia and Mississippi, according to the report by Environment America[3]. Additionally, Oklahoma, New Mexico, and Texas, where fracking is the most prevalent, have lower ALE coefficients for fracking than most states. The possible reason for this is that Fracking ALE is positively correlated with some socioeconomic factors, such as family mean income and marital rate (see Fig. 7), which are negatively correlated with stroke mortality over 65 years (see Fig. 8).

0.1 Comparative Analysis on the Effect of Fracking on Gender-based Stroke Mortality
To explore spatio-temporal differences in the effect of hydraulic fracturing on gender-based stroke mortality, we performed spatio-temporal regressions on male and female stroke mortality separately using the selected explanatory variables. Based on the regression coefficients obtained for the two groups, the temporal and spatial characteristics of different genders were

Figure 1. Temporal distribution of the average coefficients of the ALE of Fracking for 19 states with active fracking
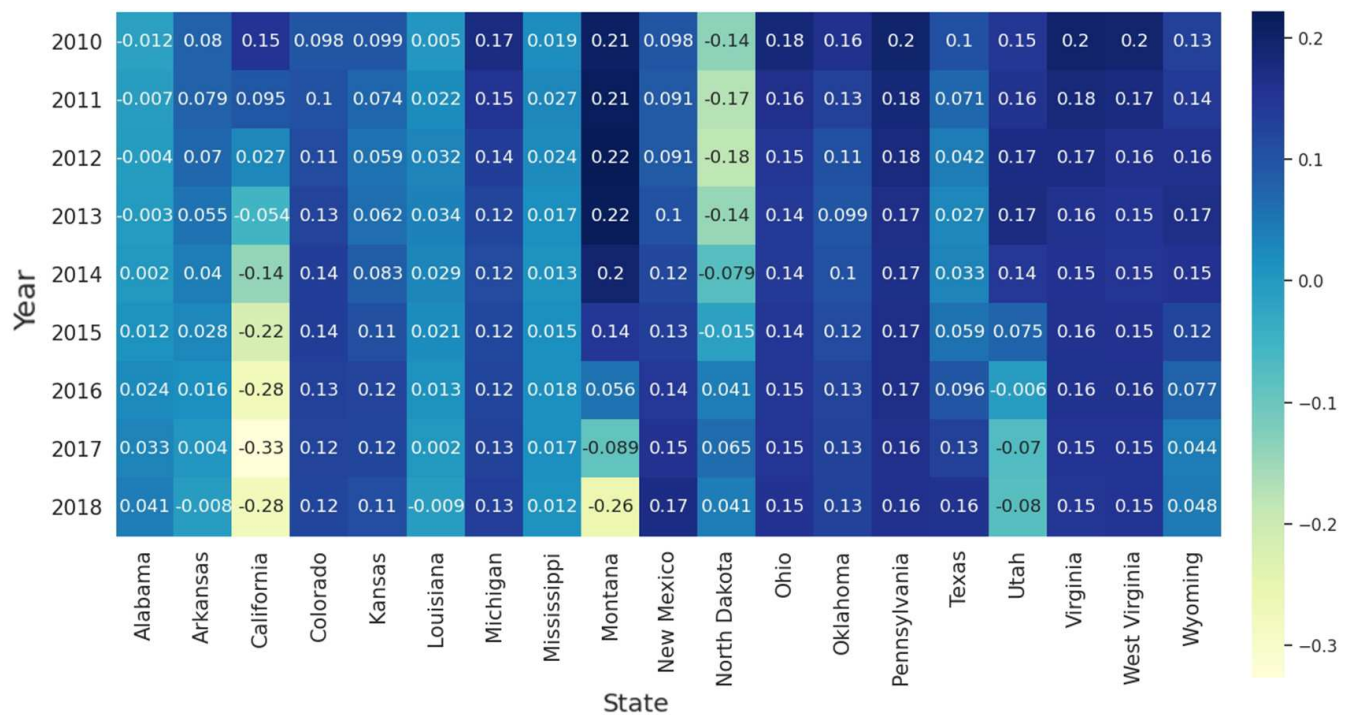
compared and analyzed.

Figs. 3 and 4 show, respectively, the regression coefficients of fracking ALE for different dependent variables (male stroke mortality and female stroke mortality). We found that the fracking ALE regression coefficients for males with stroke mortality and females with stroke mortality have similar temporal trends in Ohio, West Virginia, Virginia, and Pennsylvania. They slowly decreased from 2010 to 2018. The correlation coefficient between fracking ALE and stroke mortality (both men and women) in California decreased significantly year by year. The correlation coefficient between fracking ALE and stroke mortality (both male and female) in North Dakota increased significantly year by year. The correlation coefficients in the state of Colorado first increased and then decreased for male stroke mortality, but reversed for female stroke mortality. Additionally, the correlation coefficients between ALE from fracking and stroke mortality (both male and female) in California and some states (Arkansas, Louisiana, Mississippi, and Alabama) in the stroke belt are lower than those of other states from fracking. It demonstrates that fracking activity is not a significant factor that affects stroke mortality.

In addition to the temporal characteristics of gender-based stroke mortality, the fracking ALE regression coefficients are spatially differentiated (see Fig. 5). The ALE coefficients of fracking for male stroke mortality around Marcellus shale (including Pennsylvania, West Virginia, and Ohio) are higher than other fracking states. In contrast, the ALE coefficients for fracking for female stroke mortality in Montana, Wyoming, New Mexico, and Oklahoma are higher than those of other fracking states.

Discussion

Does Fracking Cause a Higher Risk of Stroke?

Despite the potential health risks associated with fracking, there have been several quantitative studies on how fracking can affect public health on a local scale[35,42], which characterize fracking risk as distance from the patient's residence to the nearest well. However, these methods are based on privacy-protected clinical data that contain large amounts of patient personal information. They are not applicable to this study, since patient personal information is not included in the publicly available dataset from CDC. This study uses ALE to quantitatively study the possible connection between fracking and stroke using publicly available data. It can be observed from the results of GTWR analysis that fracking has a non-negligible effect on stroke mortality over 65 in most of the areas with fracking prevalent, as shown in Tab. 2 and Fig. 2. However, its impact is relatively minor compared to other more dominant factors, such as cardiovascular disease and overdose.

Figure 2. Spatial distribution of the average coefficients of fracking ALE for 19 states related to fracking
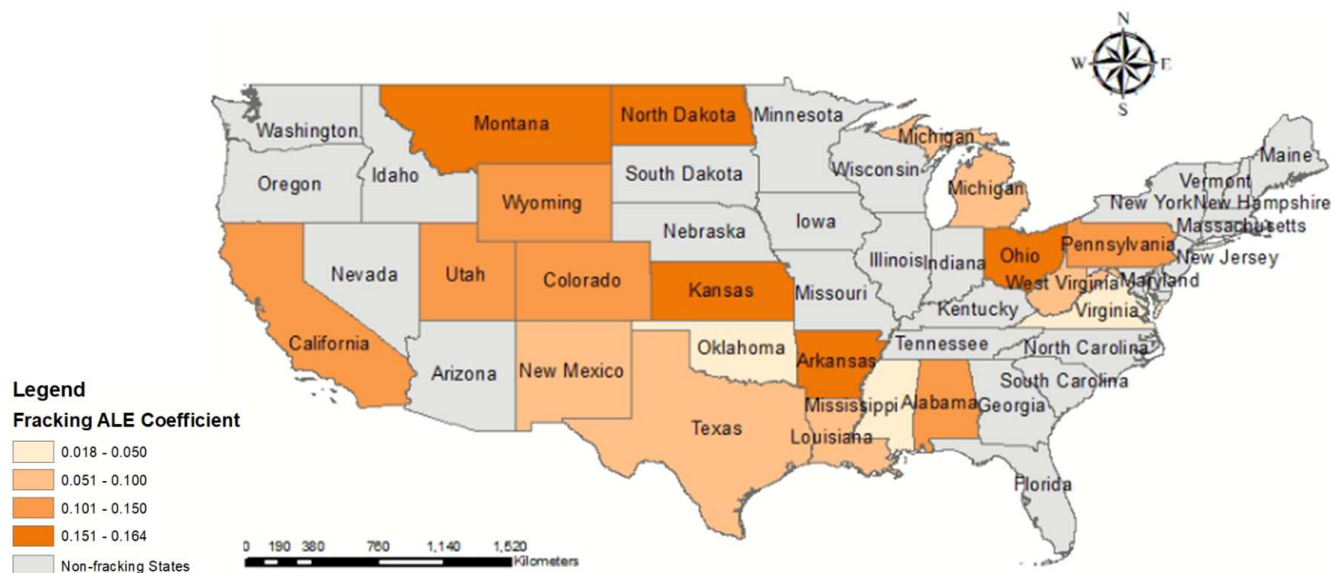
Spatiotemporal Differences in the Effect of Fracking on Age/gender-based Stroke Mortality

Deeper reasoning is needed to understand the geographic variations of fracking-related stroke mortality cases based on the age and sex of patients. As shown in Tab. 6 and 7, the Fracking ALE variable for stroke mortality between 45-64 is not considered statistically significant with a p-value > 0.1, but the Fracking ALE variable for stroke mortality over 65 is considered statistically significant with a p-value < 0.01. For gender-based stroke mortality, the higher Fracking ALE coefficient for men than for women seems to suggest that men are at higher risk, at least in regions near Marcellus shale. Additionally, higher fracking ALE coefficients for women than for men suggest that women are at greater risk in some states, including Montana, Wyoming, Oklahoma, and New Mexico. Furthermore, the fracking ALE coefficients for both male and female stroke mortality are small in some states of the stroke belt (e.g. Arkansas, Louisiana, Mississippi, and Alabama), suggesting that fracking is not a major factor in stroke mortality compared to some major factors, such as cardiovascular disease.
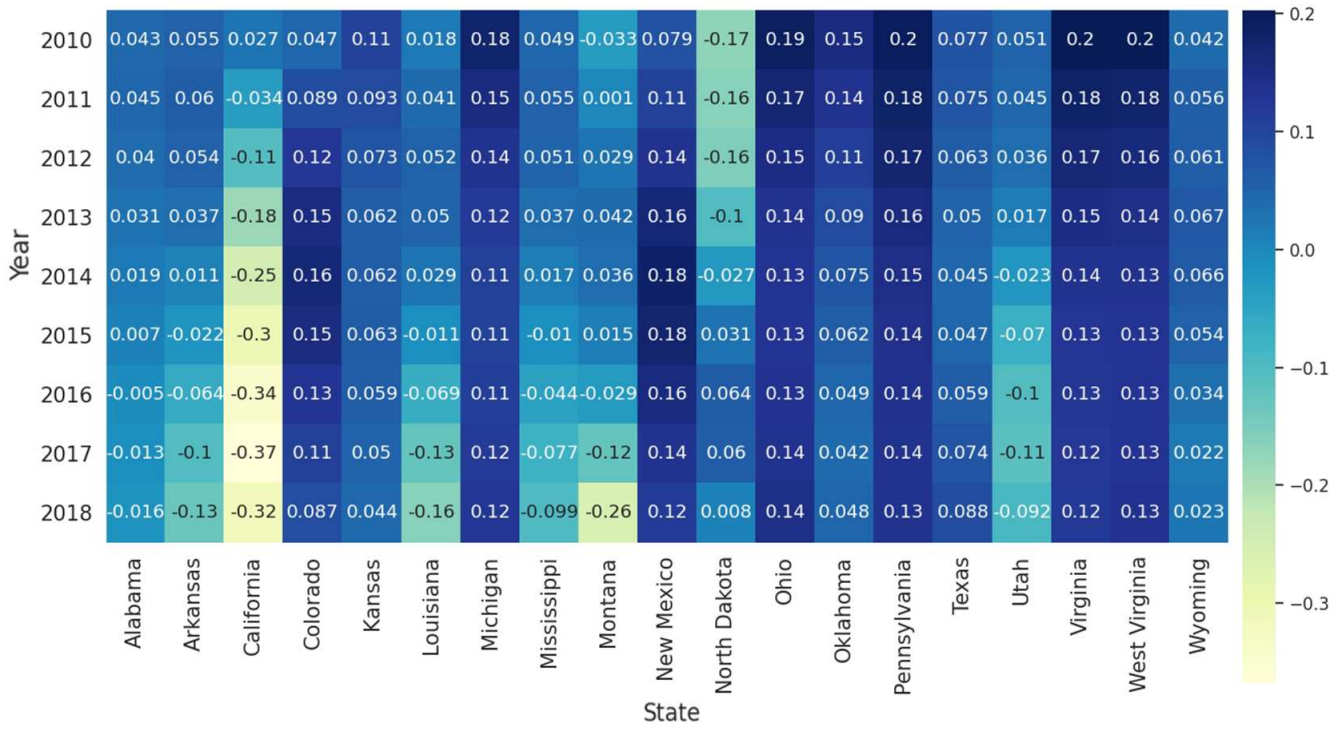
Air Pollutant Emissions from Fracking

We examined a possible mechanism of how fracking threatens public health. The existing report indicates that fracking produces environmental pollution, including hazardous water pollutants and hazardous air pollutants (HAPs)[43]. In this study, we focus on air pollutants in the state of Colorado. We have drawn a zone of interest with a radius of two kilometers around the HAP monitoring station and calculated the number of fracking wells. A HAP monitoring station in Colorado (longitude: -108.053259 and latitude: 39.453654) was selected to analyze the correlation between fracking activities and different HAPs. Four HAPs (Butadiene, Benzene, Formaldehyde, and Acetaldehyde) were monitored through this station, and we upsampled the four HAP monitoring data through linear interpolation to ensure that all four HAPs had the same time resolution. We then constructed a time series of fracking activities according to the start and end times of each fracking activity, and the overlapping fracking activities were aggregated. Finally, the four HAPs and the time series of the fracking activity were normalized, and then the fracking activities were analyzed for time series correlation with different HAPs using Pearson's correlation[44]. Pearson's r, which ranges from 0 to 1, was calculated to measure the degree of correlation between the two time series. A larger Pearson's r suggests a stronger correlation. Tab. 4 shows that the Pearson correlation (r) of benzene was higher than the other three HAPs and statistically significant (p-value < 0.01).

What is the Implication of This Study on Health Policy-making?

With the increase in fracking activity, the socioeconomic environment, such as the employment rate and family income, continues to increase. At the same time, more and more people and communities in areas with many hydraulically fractured wells report health problems, such as cancer and harm to the nervous, respiratory, and immune systems. Macroscopically, we found that the effects of fracking and stroke mortality were not significant for the 19 active states of fracking in the US. This is likely due to the following reasons: 1) Fracking areas in the US are mainly distributed in the Great Plains and Marcellus Shale, which are often located in mountainous or suburban areas with low population density, which may lead to limited spatial impact.

Figure 3. Temporal distribution of the average coefficients of the ALE of Fracking for 19 states related to fracking (Male)

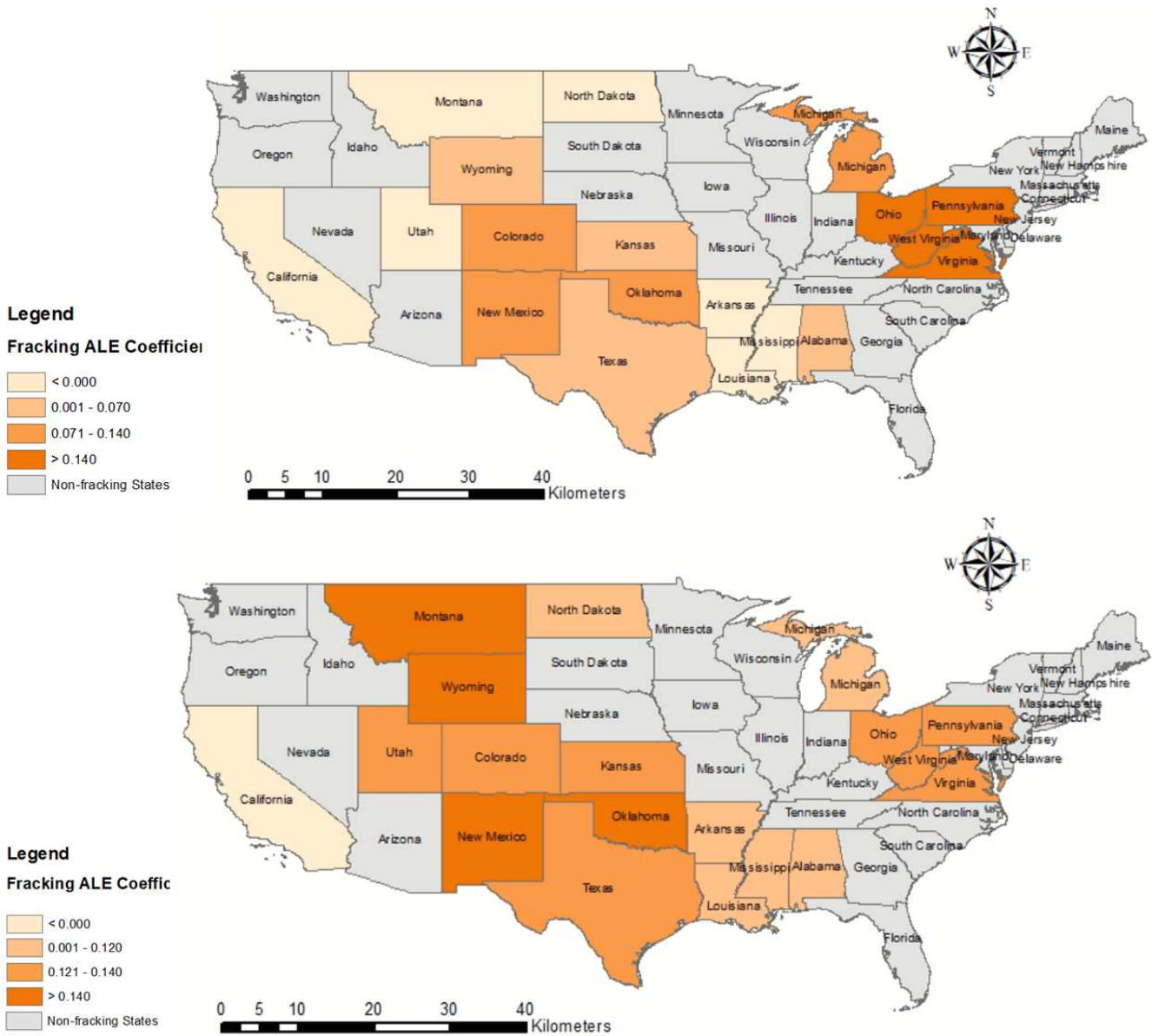| Year	Alabama	Arkansas	California	Colorado	Kansas	Louisiana	Michigan	Mississippi	Montana	New Mexico	North Dakota	Ohio	Oklahoma	Pennsylvania	Texas	Utah	Virginia	West Virginia	Wyoming
2010	0.043	0.055	0.027	0.047	0.11	0.018	0.18	0.049	-0.033	0.079	-0.17	0.19	0.15	0.2	0.077	0.051	0.2	0.2	0.042
2011	0.045	0.06	-0.034	0.089	0.093	0.041	0.15	0.055	0.001	0.11	-0.16	0.17	0.14	0.18	0.075	0.045	0.18	0.18	0.056
2012	0.04	0.054	-0.11	0.12	0.073	0.052	0.14	0.051	0.029	0.14	-0.16	0.15	0.11	0.17	0.063	0.036	0.17	0.16	0.061
2013	0.031	0.037	-0.18	0.15	0.062	0.05	0.12	0.037	0.042	0.16	-0.1	0.14	0.09	0.16	0.05	0.017	0.15	0.14	0.067
2014	0.019	0.011	-0.25	0.16	0.062	0.029	0.11	0.017	0.036	0.18	-0.027	0.13	0.075	0.15	0.045	-0.023	0.14	0.13	0.066
2015	0.007	-0.022	-0.3	0.15	0.063	-0.011	0.11	-0.01	0.015	0.18	0.031	0.13	0.062	0.14	0.047	-0.07	0.13	0.13	0.054
2016	-0.005	-0.064	-0.34	0.13	0.059	-0.069	0.11	-0.044	-0.029	0.16	0.064	0.13	0.049	0.14	0.059	-0.1	0.13	0.13	0.034
2017	-0.013	-0.1	-0.37	0.11	0.05	-0.13	0.12	-0.077	-0.12	0.14	0.06	0.14	0.042	0.14	0.074	-0.11	0.12	0.13	0.022
2018	-0.016	-0.13	-0.32	0.087	0.044	-0.16	0.12	-0.099	-0.26	0.12	0.008	0.14	0.048	0.13	0.088	-0.092	0.12	0.13	0.023

Figure 4. Temporal distribution of the average coefficients of ALE of fracking for 19 states related to fracking (Female)

Year	Alabama	Arkansas	California	Colorado	Kansas	Louisiana	Michigan	Mississippi	Montana	New Mexico	North Dakota	Ohio	Oklahoma	Pennsylvania	Texas	Utah	Virginia	West Virginia	Wyoming
2010	-0.004	0.091	0.2	0.14	0.14	-0.003	0.15	0.024	0.34	0.13	0.02	0.17	0.17	0.16	0.15	0.18	0.17	0.17	0.2
2011	0.005	0.093	0.16	0.13	0.11	0.025	0.13	0.037	0.32	0.12	-0.025	0.14	0.15	0.14	0.11	0.19	0.15	0.15	0.19
2012	0.011	0.088	0.1	0.12	0.088	0.043	0.12	0.04	0.3	0.11	-0.043	0.13	0.12	0.14	0.075	0.2	0.14	0.13	0.19
2013	0.017	0.077	0.014	0.12	0.09	0.053	0.11	0.038	0.27	0.11	-0.022	0.12	0.12	0.13	0.057	0.2	0.13	0.12	0.19
2014	0.029	0.066	-0.083	0.13	0.12	0.056	0.1	0.04	0.22	0.12	0.035	0.12	0.13	0.13	0.064	0.17	0.12	0.12	0.16
2015	0.047	0.06	-0.17	0.14	0.14	0.06	0.1	0.05	0.16	0.14	0.1	0.13	0.15	0.14	0.097	0.11	0.13	0.13	0.14
2016	0.064	0.057	-0.22	0.15	0.16	0.065	0.11	0.061	0.077	0.16	0.16	0.13	0.16	0.14	0.14	0.045	0.13	0.13	0.12
2017	0.075	0.048	-0.21	0.16	0.15	0.062	0.12	0.061	-0.015	0.19	0.18	0.14	0.16	0.13	0.18	0.038	0.12	0.13	0.13
2018	0.081	0.035	-0.11	0.16	0.14	0.052	0.12	0.054	-0.069	0.21	0.14	0.14	0.16	0.13	0.22	0.1	0.12	0.13	0.16

Figure 5. Spatial distribution of the average coefficients of fracking ALE for 19 states related to fracking for Male (upper) and Female (lower).

Table 4. Pearson correlation significance test

Air Pollutants	Pearson r	P-value
Butadiene	0.119	0.083
Formaldehyde	0.093	0.175
Acetaldehyde	0.049	0.474
Benzene	0.245	0.000 *

* indicates P < 0.01

2) The fracking process generally only lasts 3-5 days, which leads to a limited impact on time. 3) The development of fracking will promote local socioeconomic status (e.g., employment rate and marital rate), which is negatively related to stroke mortality. This may cause the effect of fracking on stroke mortality to be insignificant in some states (e.g., Texas) where fracking is active.

This study analyzed the correlation between fracking and different hazardous air pollutants based on the public air pollutant dataset from the US Environmental Protection Agency (EPA). We found that the concentration of benzene in the air was related to the calculated sequence of fracking activity in the buffer zone within a radius of 2 km (Pearson $r = 0.2452$ and the p-value $<$ 0.01). To minimize health risk, this research suggests that there should be no public facilities with a high population density within 2 km of fracking activities. Furthermore, the high concentration of benzene in the air may be due to fracking, which caused groundwater pollution due to the extremely high volatility of benzene. It may be important to identify and investigate domestic water wells that are within two kilometers of a fracking well. A more systematic study of the impact of fracking on water contamination[23] remains for future research.

Methods

Study Area and Data Collection

To study the spatio-temporal correlations of stroke mortality and fracking, the present study has chosen 49 states in the US as the study area. Alaska is not included because of its geographical isolation. Fig. 6 shows the stroke death rate per 100,000 people over 65 years of age and all the sites of fracking activity before 2018. To explore the impact of fracking on stroke mortality, we divide the 49 states into *fracking states* and *non-fracking states*. As a result, there are 24 fracking states that had fracking activities (including 19 states with active fracking and 5 states with little fracking) and 25 non-fracking states that did not have fracking activity by 2018.

Stroke is closely related to people's behavior habits (tobacco use, high-cholesterol diet, and physical activity index), socioeconomic status (family mean income, marital rate, and employment rate), and other diseases (cardiovascular, overdose, and diabetes)[28,45]. Therefore, these variables have been selected for comparison with risk factors for fracking. Tab. 5 shows details of the dependent and explanatory variables and their data sources. All data were collected in the United States from 2010 to 2018. The scale of the data we collected is state-level except for fracking, because the county-level data contains a lot of missing data that can cause problems with the analysis. Furthermore, since stroke is a chronic disease, we processed the fracking data into the cumulative number of fracking wells. In other words, the number of fracking wells in any year includes all fracking wells before that year.

Table 5. Description of the data used in this study (CDC: Centers for Disease Control and Prevention; USCB: The United States Census Bureau; FF: US FracFocus).

Type	Variables	Description	Source
Fracking	Fracking activity	The location of Fracking wells	FF
Stroke mortality	65+ stroke mortality	Stroke deaths rate over 65 per 100,000	CDC
	Male stroke mortality	Male Stroke deaths rate per 100,000	CDC
	Female stroke mortality	Female Stroke deaths rate per 100,000	CDC
Disease	Diabetes	Proportion of diagnosed diabetes	CDC
	Cardiovascular	Cardiovascular deaths rate over 65 per 100,000	CDC
	Overdose	Drug overdose death rates	CDC
	Hypertension	High blood pressure deaths over 65 per 100,000	CDC
	Obesity	Adults with a BMI > 30	CDC
Behavior	Tobacco use	Current cigarette use by adults	CDC
	High cholesterol	High total cholesterol among adults	CDC
	Physical activity index	Physical Inactivity Prevalence	CDC
	Heavy Drink	8 or more drinks per week (female) or 15 or more drinks per week (male)	CDC
Socioeconomic	Mean income	Family income by number of workers	USCB
	Marital rate	Proportion of married population	USCB
	Employment rate	Proportion of employed population	USCB
	Education	Bachelor's degree or higher	USCB

Quantitative Risk Analysis: Annual Loss Expectancy of Fracking

Recent research illustrated that 2-16% of oil and gas wells spill liquids every year[46]. These accidents have caused fracking chemicals to contaminate drinking water and air, further raising serious public health risks[47–49]. To quantify the public health risk of fracking, the annualized loss expectancy (ALE) model[50] has been applied to assess public health-related ALE caused by fracking by state, which is abbreviated as Fracking ALE. ALE is the product of the annual rate of occurrence (ARO) and the single loss expectancy (SLE)[51] caused by fracking, as shown in Eq. (1).

$$ALE = ARO \times SLE \tag{1}$$

where fracking ARO is the annual accident occurrence of pollutant leakage due to fracking, which is equal to the density of fracking per square kilometer (FD) multiplied by the annual rate of occurrence per fracking per square kilometer (AROF), $ARO = FD \times AROF$.

Fracking SLE is defined as population with negative health impacts expected from the occurrence of fracking accident. For risk calculation, SLE is used to calculate a single loss when a specific event occurs. Fracking SLE is calculated by multiplying the Population Density per square kilometer (PD) by the fracking exposure factor (EF), $SLE = PD \times EF$, as the higher the population density in active fracking states, the greater the negative impact of fracking on public health.

Fracking ALE represents the product of AROF, FD, PD, and EF, which means the average loss per year of environmental pollution caused by fracking on public health. The FD and PD of the jth state in the ith year are FD_{ij} and PD_{ij}, respectively. As an important explanatory variable for the regression in Section 0.1.1, ALE is normalized as follows:

$$Normalized\ ALE_{ij} = \frac{ALE_{ij} - Min(ALE)}{Max(ALE) - Min(ALE)} \tag{2}$$

where $Normalized\ ALE_{ij}$ is the normalized ALE in state j in year i. Since AROF and EF are constants, Equation 2 can be further formulated as

$$Normalized\ ALE_{ij} = \frac{FD_{ij} \times PD_{ij} - Min(FD \times PD)}{Max(FD \times PD) - Min(FD \times PD)} \tag{3}$$

Regression model

Multicollinearity

We performed multicollinearity diagnostics before applying the regression model, since several explanatory variables could be highly correlated. Multicollinearity means that there is a high linear correlation between several specific explanatory variables, which could lead to bias in explaining the significance and associations of other variables. We adopted the variance inflation factor (VIF)[52], a metric of the severity of multicollinearity, to eliminate this problem. Explanatory variables with a VIF greater than 10 are considered to cause multicollinearity and should be excluded from the model[53].

0.1.1 Spatial Autocorrelation

As a commonly used spatial autocorrelation test, Moran's I test represents the spatial autocorrelation of a single explanatory variable and can be expressed as[54].

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

Moran's I ranges from -1 to 1. A higher positive value indicates that closer observations have more similar attribute values, whereas farther observations have more distinct attribute values, indicating spatial aggregation. Negative values represent a spatially dispersed distribution, and a zero value represents a spatially random distribution. The null hypothesis of the Moran's I test indicates that the explanatory variables are spatially independent. It indicates that the Moran's I is close to zero. The $Z-$score is used as a significant indicator to measure Moran's I to verify the null hypothesis, whose formula is as follows[54].

$$Z(I) = \frac{I - E(I)}{\sqrt{Var(I)}} \tag{5}$$

where $E(I)$ and $Var(I)$ are the expectation and standard deviation of Moran's I, respectively. The significance level in this study is established as the p-value < 0.05.

Geographical and Temporal Weighted Regression (GTWR)

To analyze the spatio-temporal correlation between risk factors (see Tab. 5) and stroke mortality, the GTWR model was selected as the regression model. Compared to traditional Geographically Weighted Regression (GWR)[36], which only considers spatial features, the GTWR model[37] considers the non-stationary effect in space and time. Therefore, it was adopted to explore the spatio-temporal heterogeneity of the influence of fracking on different stroke mortality (i.e., age-based and gender-based) under the constraint of spatio-temporal differences. By establishing a three-dimensional elliptical coordinate system (including

time, longitude, and latitude) in which the temporal dimension is the vertical dimension, in addition to the two horizontal spatial dimensions (longitude and latitude), the model can describe the spatio-temporal influence via the regression coefficients corresponding to explanatory variables. The GTWR model is described as

$$\mathbf{Y}_i = \beta_0(u_i, v_i, t_i) + \sum_k \beta_k(u_i, v_i, t_i) X_{ik} + \varepsilon_i \tag{6}$$

where u and v represent longitude and latitude, respectively. $\beta_0(u_i, v_i, t_i)$ represents the intercept item of state with centroid at (u_i, v_i) in year t_i; k is the number of explanatory variables; $\beta_k(u_i, v_i, t_i)$ is the regression coefficient of the kth explanatory variable in year t; X_{ik} is the kth explanatory variable. These explanatory variables have been defined in the Tab. 5.

Correctional values of the Akaike Information Criterion (AICc)[55] are an important metric and are often used to select explanatory variables and determine the final model with the lowest AICc. In this study, the ArcGIS GTWR plugin was used to analyze the spatio-temporal correlation between stroke mortality and risk factors such as fracking ALE.

Conclusion

This article provided a systematic study on the spatiotemporal correlation between fracking and stroke mortality using the GTWR model. The temporal trend of positive correlation between fracking ALE and stroke mortality shows a varying pattern from state to state. The spatial distribution appears to demonstrate that there is a gender difference between the Great Plains and the Marcellus Shale. Our regression results also show that disease-related risk factors, including cardiovascular and overdose, have a more significant correlation with stroke mortality over 65 than those related to fracking. Finally, there appears to be a significant temporal dependency between fracking and air pollutant emissions, especially for benzene. Future studies Future studies may focus on developing county-level GWR/GTWR models, although missing data is a critical challenge to be resolved.

Data Availability Statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Gandossi, L. & Von Estorff, U. An overview of hydraulic fracturing and other formation stimulation technologies for shale gas production. *Eur. Commisison Jt. Res. Cent. Tech. Reports* **26347** (2013).

2. Montgomery, C. T. & Smith, M. B. Hydraulic fracturing: History of an enduring technology. *J. Petroleum Technol.* **62**, 26–40 (2010).

3. Administration, U. E. I. Hydraulic fracturing accounts for about half of current u.s. crude oil production (2016).

4. Finkel, M. L. & Hays, J. The implications of unconventional drilling for natural gas: a global public health concern. *Public health* **127**, 889–893 (2013).

5. Meng, Q. Characterizing and modeling environmental emergency of unconventional oil and gas spills in the USA: Life-year versus spill factors. *J. Clean. Prod.* **237**, 117794 (2019).

6. Kovats, S. *et al.* The health implications of fracking. *The Lancet* **383**, 757 (2014).

7. Black, K. J., Boslett, A. J., Hill, E. L., Ma, L. & McCoy, S. J. Economic, environmental, and health impacts of the fracking boom. *Annu. Rev. Resour. Econ.* **13**, 311–334 (2021).

8. Soyer, M., Kaminski, K. & Ziyanak, S. Socio-psychological impacts of hydraulic fracturing on community health and well-being. *Int. J. Environ. Res. Public Heal.* **17**, 1186 (2020).

9. Caron-Beaudoin, É. *et al.* Volatile organic compounds (vocs) in indoor air and tap water samples in residences of pregnant women living in an area of unconventional natural gas operations: Findings from the experiva study. *Sci. The Total. Environ.* **805**, 150242 (2022).

10. Ma, L., Hurtado, A., Eguilior, S. & Borrajo, J. F. L. Exposure risk assessment to organic compounds based on their concentrations in return water from shale gas developments. *Sci. The Total. Environ.* 153586 (2022).

11. Meng, Q. The impacts of fracking on the environment: A total environmental study paradigm. *Sci. The Total. Environ.* **580**, 953–957 (2017).

12. McKenzie, L. M., Witter, R. Z., Newman, L. S. & Adgate, J. L. Human health risk assessment of air emissions from development of unconventional natural gas resources. *Sci. Total. Environ.* **424**, 79–87 (2012).

13. Vengosh, A., Jackson, R. B., Warner, N., Darrah, T. H. & Kondash, A. A critical review of the risks to water resources from unconventional shale gas development and hydraulic fracturing in the united states. *Environ. science & technology* **48**, 8334–8348 (2014).

14. Deziel, N. C. *et al.* Unconventional oil and gas development and health outcomes: A scoping review of the epidemiological research. *Environ. Res.* **182**, 109124 (2020).

15. Tao, Z., Liu, C., He, Q., Chang, H. & Ma, J. Detection and treatment of organic matters in hydraulic fracturing wastewater from shale gas extraction: A critical review. *Sci. The Total. Environ.* 153887 (2022).

16. Meng, Q. Rethink potential risks of toxic emissions from natural gas and oil mining. *Environ. Pollut.* **240**, 848–857 (2018).

17. Casey, J. A. *et al.* Unconventional natural gas development and birth outcomes in pennsylvania, usa. *Epidemiol. (Cambridge, Mass.)* **27**, 163 (2016).

18. Rasmussen, S. G. *et al.* Association between unconventional natural gas development in the marcellus shale and asthma exacerbations. *JAMA internal medicine* **176**, 1334–1343 (2016).

19. Tustin, A. W. *et al.* Associations between unconventional natural gas development and nasal and sinus, migraine headache, and fatigue symptoms in pennsylvania. *Environ. health perspectives* **125**, 189–197 (2017).

20. Casey, J. A., Wilcox, H. C., Hirsch, A. G., Pollak, J. & Schwartz, B. S. Associations of unconventional natural gas development with depression symptoms and disordered sleep in pennsylvania. *Sci. reports* **8**, 1–10 (2018).

21. Marlin-Tackie, F. A. & Smith, J. M. Key characteristics influencing risk perceptions of unconventional energy development. *J. Clean. Prod.* **251**, 119644 (2020).

22. McDermott-Levy, R., Kaktins, N. & Sattler, B. Fracking, the environment, and health. *AJN The Am. J. Nurs.* **113**, 45–51 (2013).

23. Jackson, R. B. *et al.* The environmental costs and benefits of fracking. *Annu. review Environ. Resour.* **39**, 327–362 (2014).

24. Stacy, S. L. A review of the human health impacts of unconventional natural gas development. *Curr. epidemiology reports* **4**, 38–45 (2017).

25. Towfighi, A. & Saver, J. L. Stroke declines from third to fourth leading cause of death in the united states: historical perspective and challenges ahead. *Stroke* **42**, 2351–2355 (2011).

26. Lanska, D. J. & Kuller, L. H. The geography of stroke mortality in the united states and the concept of a stroke belt. *Stroke* **26**, 1145–1149 (1995).

27. Liao, Y., Greenlund, K. J., Croft, J. B., Keenan, N. L. & Giles, W. H. Factors explaining excess stroke prevalence in the us stroke belt. *Stroke* **40**, 3336–3341 (2009).

28. Feigin, V. L. *et al.* Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the global burden of disease study 2013. *The Lancet Neurol.* **15**, 913–924 (2016).

29. Howard, G. & Howard, V. J. Twenty years of progress toward understanding the stroke belt. *Stroke* **51**, 742–750 (2020).

30. Srebotnjak, T. & Rotkin-Ellman, M. Fracking fumes: Air pollution from hydraulic fracturing threatens public health and communities. *Nat. Resour. Def. Counc.* (2014).

31. Law, A., Duff, D., Saunders, P., Middleton, J. & McCoy, D. Medical organisations must divest from fossil fuels (2018).

32. Krajnak, K. *et al.* Biological effects of inhaled hydraulic fracturing sand dust. vi. cardiovascular effects. *Toxicol. applied pharmacology* **406**, 115242 (2020).

33. Oikonomou, P. D. *et al.* Water acquisition and use during unconventional oil and gas development and the existing data challenges: Weld and garfield counties, co. *J. environmental management* **181**, 36–47 (2016).

34. Purvis, R. M. *et al.* Effects of 'pre-fracking'operations on ambient air quality at a shale gas exploration site in rural north yorkshire, england. *Sci. total environment* **673**, 445–454 (2019).

35. Currie, J., Greenstone, M. & Meckel, K. Hydraulic fracturing and infant health: New evidence from pennsylvania. *Sci. advances* **3**, e1603021 (2017).

36. Brunsdon, C., Fotheringham, A. S. & Charlton, M. E. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geogr. analysis* **28**, 281–298 (1996).

37. Huang, B., Wu, B. & Barry, M. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *Int. J. Geogr. Inf. Sci.* **24**, 383–401 (2010).

38. Dismuke, C. & Lindrooth, R. Ordinary least squares. *Methods Des. for Outcomes Res.* **93**, 93–104 (2006).

39. Arboix, A. Cardiovascular risk factors for acute stroke: Risk profiles in the different subtypes of ischemic stroke. *World J. Clin. Cases: WJCC* **3**, 418 (2015).

40. Avan, A. *et al.* Socioeconomic status and stroke incidence, prevalence, mortality, and worldwide burden: an ecological analysis from the global burden of disease study 2017. *BMC medicine* **17**, 1–30 (2019).

41. Center, E. A. R. . P. Fracking by the numbers the damage to our water, land and climate from a decade of dirty drilling (2016).

42. Clark, C. J. *et al.* Assessing unconventional oil and gas exposure in the appalachian basin: comparison of exposure surrogates and residential drinking water measurements. *Environ. science & technology* (2022).

43. Windham, G. C., Zhang, L., Gunier, R., Croen, L. A. & Grether, J. K. Autism spectrum disorders in relation to distribution of hazardous air pollutants in the san francisco bay area. *Environ. health perspectives* **114**, 1438–1444 (2006).

44. Benesty, J., Chen, J., Huang, Y. & Cohen, I. Pearson correlation coefficient. In *Noise reduction in speech processing*, 1–4 (Springer, 2009).

45. Feigin, V. L. *et al.* Global and regional burden of stroke during 1990–2010: findings from the global burden of disease study 2010. *The lancet* **383**, 245–255 (2014).

46. Patterson, L. A. *et al.* Unconventional oil and gas spills: risks, mitigation priorities, and state reporting requirements. *Environ. Sci. & Technol.* **51**, 2563–2573 (2017).

47. Korfmacher, K. S., Jones, W. A., Malone, S. L. & Vinci, L. F. Public health and high volume hydraulic fracturing. *NEW SOLUTIONS: A J. Environ. Occup. Heal. Policy* **23**, 13–31 (2013).

48. Colborn, T., Kwiatkowski, C., Schultz, K. & Bachran, M. Natural gas operations from a public health perspective. *Hum. ecological risk assessment: An Int. J.* **17**, 1039–1056 (2011).

49. Bamberger, M. & Oswald, R. E. Impacts of gas drilling on human and animal health. *New solutions: a journal environmental occupational health policy* **22**, 51–77 (2012).

50. Bistarelli, S., Fioravanti, F. & Peretti, P. Defense trees for economic evaluation of security investments. In *First International Conference on Availability, Reliability and Security (ARES'06)*, 8–pp (IEEE, 2006).

51. Karabacak, B. & Sogukpinar, I. Isram: information security risk analysis method. *Comput. & Secur.* **24**, 147–159 (2005).

52. Akinwande, M. O., Dikko, H. G., Samson, A. *et al.* Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis. *Open J. Stat.* **5**, 754 (2015).

53. Kutner, M. H., Nachtsheim, C. J., Neter, J. & Wasserman, W. *Applied linear regression models*, vol. 4 (McGraw-Hill/Irwin New York, 2004).

54. Moran, P. A. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).

55. Cavanaugh, J. E. Unifying the derivations for the akaike and corrected akaike information criteria. *Stat. & Probab. Lett.* **33**, 201–208 (1997).

Supplementary Files

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

- ScientificReportssupp.pdf