JOURNAL OF
EMERGING INVESTIGATORS

Article

Understanding the correlation between various pollutants and cancer across geographical clusters in the U.S.

Atreya Naik[1], Morteza Sarmadi[2]
[1] The Overlake School, Redmond, Washington
[2] Alumni, Massachusetts Institute of Technology, Cambridge, Massachusetts

SUMMARY
Air, water, and chemical pollutants harm human health by damaging vital organs, causing gastrointestinal and neurological issues, and contributing to cancer and immune system dysfunction. Many prior studies have shown that the risk of obtaining life-threatening diseases in humans increases when exposed to high concentrations of pollutants. Our study explored whether different types of pollutants are positively correlated with cancer incidence rates. We hypothesized that there is a statistically significant positive correlation between environmental pollutants and cancer incidence rates across different geographical clusters of the United States. We used the Pearson correlation coefficient to measure the strength of the correlation across three years (2018–2020). To assess statistical significance, we calculated a t-statistic for each coefficient, with $p < 0.05$ considered statistically significant. Thirty-six different types of cancer and twelve pollutant metrics were analyzed using publicly available data from the Centers for Disease Control and Prevention (CDC) and Environmental Protection Agency (EPA). Our research showed that certain clusters in the United States showcase a strong correlation between different environmental pollutants and various cancers. These findings can potentially help prevent premature deaths caused by cancer and help explain why certain states in the United States have larger cancer incidence rates than others.

INTRODUCTION
In the United States, cancer is one of the primary causes of death (1). From 2016 to 2020, there were over eight million cases of cancer. In the same timeframe, over three million people died of cancer-associated deaths (2). Furthermore, between 2014 and 2018, the cancer incidence rates for the age groups between 15-39 years rose by an average of almost 1% per year (3). People residing in the United States are, on average, first diagnosed with cancer at 66 years, around 11 years under the average life expectancy (4).

Pollution is defined as the introduction of harmful substances into the environment and represents a significant public health concern due to its wide-ranging effects on human well-being (5). Environmental pollutants can be broadly categorized into air pollutants, water pollutants, and chemical pollutants. Air pollutants, such as particulate matter and ozone, are associated with respiratory and cardiovascular diseases; water pollutants can compromise drinking water quality and food safety; and chemical pollutants, including lead and benzene, released through industrial activities, may persist in soil and ecosystems, disrupting biological systems and limiting access to safe water and agricultural resources. As a result, pollutant exposure can adversely affect human health not only by increasing disease risk but also by indirectly impacting essential resources such as clean air, potable water, and food supplies.

Numerous environmental epidemiological studies have shown that exposure to air, water, and chemical pollutants is associated with increased risk of multiple cancer types due to mechanisms such as chronic inflammation, oxidative stress, and DNA damage (6 - 8). For example, long-term exposure to fine particulate matter specifically PM2.5, a major air pollutant, has been associated with an increased risk of lung cancer (6). Similarly, nitrate contamination in drinking water has been linked to an elevated risk of colorectal cancer through the formation of carcinogenic N-nitroso compounds (7). In addition, exposure to chemical pollutants such as benzene has been associated with an increased risk of leukemia and other hematologic malignancies (8). According to the American Lung Association, in the United States, approximately 156 million people reside in places containing unhealthy amounts of pollution (9). Air, chemical, and water pollution remain significant environmental health concerns. National Oceanic and Atmospheric Administration (NOAA) measurements have documented continued increases in atmospheric greenhouse gas concentrations, while recent studies have highlighted the growing prevalence and diversity of industrial chemical pollutants and the widespread presence of contaminants in drinking water supplies, underscoring the potential public health impacts of long-term environmental exposure (10 - 12).

Prior studies have established a correlation between geographical location and cancer incidence. For example, residents living in Kentucky are 2.4 times as likely to be affected by lung cancer compared to those living in Utah (13). This geographic disparity may be partially explained by environmental factors, including coal-mining–related exposures in Appalachia, a region that encompasses parts of eastern Kentucky and has historically experienced elevated levels of environmental pollution (14). Geographic variation in cancer incidence has also been observed for cancers beyond lung cancer. For example, Iowa has a colorectal cancer incidence rate of approximately 42.3 cases per 100,000 individuals, compared with 29.7 cases per 100,000 in Utah, representing a substantially higher disease burden in Iowa (15). This geographic disparity may be partially explained by differences in environmental exposures, particularly nitrate contamination of groundwater and drinking water supplies as nitrate concentrations are often highest in heavily

https://doi.org/10.59720/24-322

agricultural regions such as the Midwest and High Plains because intensive fertilizer use increases nitrate runoff into groundwater and surface water systems (16). As a result, residents of these regions may experience greater exposure to a known correlate of colorectal cancer than residents of less agriculturally intensive regions (17). Similarly, geographic differences in cancer incidence have been observed between heavily industrialized regions and less industrialized areas of the United States (18). Studies have reported elevated cancer burdens in communities located along the Gulf Coast petrochemical corridor compared with many other regions of the country (18). This geographic disparity may be partially explained by higher concentrations of industrial chemical pollutants, including benzene and ethylene oxide, released from the dense network of refineries and chemical manufacturing facilities located throughout the region (19). Because exposure to benzene, ethylene oxide, and other industrial pollutants has been associated with increased cancer risk, regional differences in industrial activity may contribute to geographic variation in cancer incidence (19). Together, these examples demonstrate that geographic differences in cancer incidence often coincide with regional variation in air, water, and chemical pollution, suggesting that environmental exposures may partially explain why certain cancers occur more frequently in some regions of the United States than in others.

Given these documented associations between environmental pollutants and cancer incidence, we sought to quantitatively evaluate whether regional differences in pollutant abundance across the United States are associated with variation in cancer incidence rates. We hypothesized that differences in the abundance of environmental pollutants across geographical regions of the United States of America may influence the incidence rates of specific cancers as determined using the Pearson coefficient.

Our analysis identified strong and statistically significant correlations between environmental pollutant levels and cancer incidence rates across multiple regions of the United States. These associations were observed for different categories of pollutants, including air, water, and chemical pollutants, and varied by geographic region.

RESULTS

The aim of our study was to understand the correlation between various pollutants and various types of cancer across different regions of the United States. The eight regions considered with the different cancer and pollutant types are showcased (**Figure 1**). We analyzed 36 different

Cancer types that had incidence data:	**Pollutants considered**
acute lymphocytic, acute myeloid, brain and other nervous system, cervix, colon and rectum, corpus and uterus, cranial nerves and other nervous system, digestive system, endocrine system, esophagus, eye and orbit, female breast, gallbladder, hodgkin lymphoma, kaposi sarcoma, kidney and renal pelvis, larynx, leukemias, liver and intrahepatic bile duct, lung and bronchus, melanomas of the skin, mesothelioma, myeloma, nasopharynx, non-hodgkin lymphoma, oral cavity and pharynx, oropharynx, ovary, pancreas, prostate, respiratory system, salivary gland, stomach, testis, thyroid, urinary bladder	(For more details on what they mean, hazardous levels etc. see **Table 1**): **Air pollutants:** CO, NO2, ozone, PM2.5, PM10 **Water pollutants:** total pollutant pounds (lb./yr) for majors, total pollutant pounds (lb./yr) for non-majors **Chemical pollutants:** lead, sulfuric acid, chromium compounds, benzene, polycyclic aromatic compounds

Figure 1: Regional classification of U.S. states and variables included in the analysis. Map (35, Legends of America) showing how the different states were classified into regions, which we used for the study. All cancer types and pollutants considered in the study listed.

https://doi.org/10.59720/24-322

Pollutant	Name	Full form	Per day threshold considered harmful	Reason for choosing pollutant
Air	Days CO	Carbon monoxide	8-hour average CO concentration exceeds 9 ppm (parts per million) or 1-hour average CO exceeds 35 ppm (very rare)	Interferes with the blood's ability to carry oxygen, leading to fatigue, and in some cases death (36).
Air	Days NO2	Nitrogen dioxide	1-hour average NO2 concentration exceeds 100 ppb (parts per billion)	Inhaled through nose and mouth and penetrates lung area, causing inflammation (37).
Air	Days O3	Ozone	8-hour average O3 exceeds 70 ppb	Inhaled and absorbed deep into the lungs, causing chest tightness and coughing (38).
Air	Days PM2.5	Particulate matter with diameter < 2.5 micrometers (μm)	when 24-hour average concentration exceeds 35.5 μg/m³	Small enough to reach deep into the lungs, and can enter the bloodstream, and is commonly associated with premature mortality (lung cancer, heart attacks) (39).
Air	Days PM10	Particulate matter with diameter ≥ 2.5 μm and < 10 μm	when 24-hour average concentration exceeds 155 μg/m³	Inhaled through upper airways, leading to coughing, and exposure can harm lung function (39).
Water	Total pollutant pounds (lb./yr) for majors	Total amount of pollutants discharged or released for individual National Pollutant Discharge Elimination System (NPDES) permits with major destinations	N/A	Major facilities account for the bulk of water pollution, so this measure captures large, sustained pollutant flows into water systems. By focusing on majors, we ensure the data reflects high impact sources that can drive regional water quality trends (40).
Water	Total pollutant pounds (lb./yr) for non-majors	Total amount of pollutants discharged or released for individual NPDES permits with minor destinations	N/A	Non-majors, though smaller in scale, are more numerous and collectively significant. Tracking this metric helps identify diffuse sources of pollution that may accumulate over time, ensuring a more comprehensive view of regional water quality risks (40).
Chemical	Lead	Lead	0.15 μg/m³	Causes developmental delays, behavioral issues, and lowered IQ (41).
Chemical	Sulfuric acid	Sulfuric acid	25,000 lb manufactured/use; 10,000 lb otherwise use	It causes severe respiratory irritation when inhaled (42).
Chemical	Chromium compounds	Chromium compounds	25,000 lb/year Toxics Release Inventory (TRI) threshold	Known human carcinogen, damaging DNA and causing lung, nose, and throat cancer (43).
Chemical	Benzene	Benzene	0.005 mg/L water; 1 ppm	It causes immune system suppression and developmental effects (44).
Chemical	Polycyclic aromatic compounds	Polycyclic aromatic compounds – byproducts of incomplete combustion of organic materials	25,000 lb/year TRI threshold; ~1–10 ng/L	Contains potent carcinogens within the group of over 100 chemicals, which link to many diseases (45).

Table 1: Air, water, and chemical pollutants. The various pollutants that were considered for analysis in this study (obtained from the Environmental Protection Agency (EPA).

https://doi.org/10.59720/24-322

types of cancer obtained from the Centers for Disease Control and Prevention (CDC) database and 12 pollutant metrics obtained from Environmental Protection Agency (EPA) databases, including State Statistics database , the Toxic Release Inventory (TRI) database, and AirData, across eight regions over the years 2018, 2019, and 2020 (**Table 1**) (20 – 23). To clean and analyze the data, we used a custom Python (24) script and followed a data processing workflow (**Figure 2A-C**). After normalizing and merging cancer and environmental data, we used the Pearson Correlation Coefficient matrix (**Figure 4**) to determine the strength of the correlation between the prevalence of cancer incidence rates and air pollutants.

The findings revealed strong correlations ($r > 0.7$, $p < 0.05$) between specific air, water, and chemical pollutants and various cancers in certain regions (**Figure 3A-F**). For example, in the Midwest, across the years 2018-2020, we observed a strong correlation between air pollution (PM2.5) and skin melanoma prevalence (mean $r = 0.887$, $p = 0.0017$) (**Figure 3A, Figure 5**). In the Mid-Atlantic, chemical pollution (benzene) showed a strong correlation with colon and rectum cancer (mean $r = 0.783$, $p = 0.0328$), confirming significance (**Figure 3B**). In the Great Plains, we observed a strong correlation between water pollution and kidney and renal pelvis cancer (**Figure 3C,** mean $r = 0.717$, $p = 0.0864$).

There are other results with similar positive correlations: In the Great Plains, benzene was strongly correlated with kidney and renal pelvis cancer (mean $r = 0.867$, $p = 0.0285$) (**Table 2**). In the Southwest, PM10 (particulate matter suspended in the air with a diameter of 10 micrometers or less) had a strong correlation with esophagus cancer (mean $r = 0.913$, $p = 0.0435$), and in the South, ozone showed a significant correlation with Hodgkin lymphoma (mean $r = 0.6$, $p = 0.0196$) (**Table 2**). The West Coast showed a strong correlation between PM2.5 and skin melanomas (mean $r = 0.953$, $p = 0.0980$), and in New England, CO was significantly correlated with myeloma (mean $r = 0.80$, $p = 0.0280$) (**Table 2**). In the Rocky Mountains, leukemia showed a strong and significant correlation with PM2.5 (mean $r = 0.863$, $p = 0.0298$) (**Table 2**). These statistically supported correlations were consistent across the three-year period, reinforcing our hypothesis that pollutant levels are meaningfully linked to specific cancer incidence rates. Plotting the cancer incidence across different regions of the United States, we see that the environmental pollutants had a strong correlation to various cancers in certain regions (**Figure 6**).

DISCUSSION

Our results reveal a strong correlation between acute exposure to air, water, and chemical pollutants and increased cancer incidence in certain geographical regions of the United States (**Figure 6)**. Although this does not confirm a causal relationship, these results highlight broader geographic patterns where pollutant concentrations and cancer incidence rates are correlated.

Although it is unclear why certain air, water, and chemical pollutants have extremely high correlations to specific cancer incidence rates in specific regions and have much lower correlations in other regions, many studies have proposed broader explanations for why these pollutants have a positive correlation with cancer incidence rate (25-30) According to the National Cancer Institute, women who were exposed to higher nitrate levels had increased risks of ovarian cancer (25). We also found this specific trend across the years 2018-2020 when observing the correlation between ovarian cancer and total water pollutant pounds in certain regions (**Figure 3C**). In general, when people drink water contaminated with nitrates or arsenic, it causes the body to form certain compounds that are potent and harmful. Specifically, ingested nitrates can be reduced to nitrites and subsequently form N-nitroso compounds, including nitrosamines and nitrosamides, which are known carcinogens (26). Similarly, inorganic arsenic is metabolized into monomethylarsonic acid and dimethylarsinic acid, a process that generates reactive oxygen species and contributes to DNA damage and increased cancer risk (27).

Furthermore, a study in 2021 concluded that exposure to chemical pesticides increases the risk of brain cancer (28). We also found that chemical pollutants like benzene, lead, and sulfuric acid had a strong correlation to brain cancer in the Mid-Atlantic across the years 2018-2020 (**Figure 3E**). The general consensus is that DNA damage in somatic cells due to exposure to chemical pollutants is a primary mechanism that may lead to cancer. However, this is not the only pathway; in some cases, pollutants may instead impair the body's ability to repair existing DNA damage, contributing to cancer development indirectly (29).

According to a recent scoping review air pollution has a strong link to cervical cancer (30) also found an extremely strong correlation between cervical cancer and the air pollutant, PM2.5 in the southwest region across the years 2018-2020 (**Figure 3D**). In general, air pollutants have a strong correlation with cancer. While this is plausible, other mechanisms may also contribute such as chronic inflammation or impaired immune responses caused by prolonged pollutant exposure (30).

To account for differences in population size among states, we normalized cancer incidence using population-adjusted rates. This helped ensure that correlations were not simply a reflection of larger populations having more cancer cases. However, this does not eliminate all bias such as data limitations, regional reporting inconsistencies, and the granularity of the datasets still present challenges. Additionally, Hawaii and Alaska were excluded due to insufficient comparative data within their geographic regions.

There were many challenges faced in this case study. We chose EPA and CDC for our datasets as they had the most comprehensive data over many years. In this raw data, there were many discrepancies in the raw data when joining the cancer and environmental datasets which required substantial data cleaning and processing. This study was performed at a state granularity rather than at a city granularity because there was limited information in publicly available datasets on both cancer and environmental metrics for cities and counties. Although state-level aggregation enabled a more comprehensive and consistent analysis across regions, it represents a limitation because it can mask localized pollution hotspots and cancer clusters within states. City or county-level data would have allowed for higher spatial resolution, enabling the identification of more localized associations between specific pollutant exposures and cancer incidence.

https://doi.org/10.59720/24-322

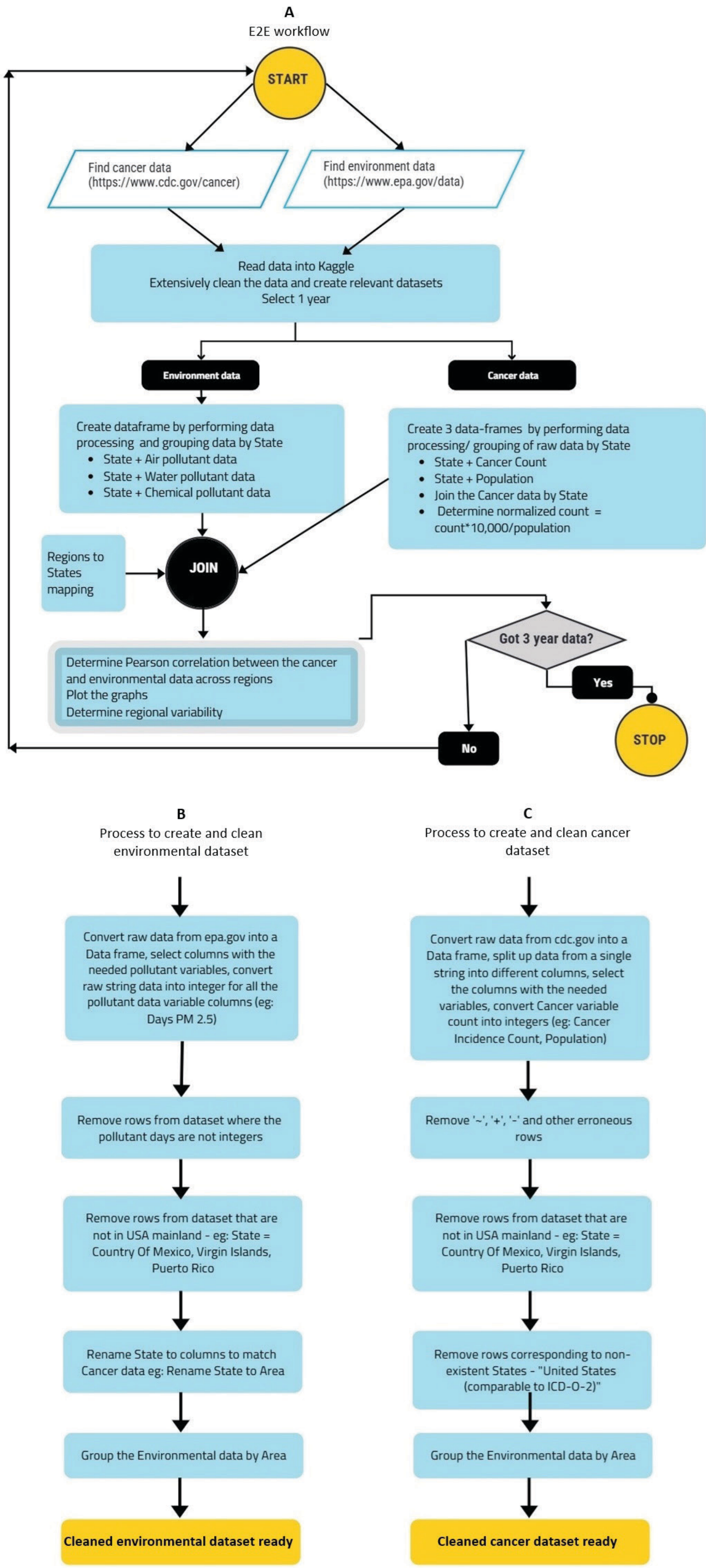


Figure 2: Schematic representation of the data processing workflow. Steps describing the process used in the study to first clean and convert raw data into usable environmental and cancer datasets, and then subsequently use those datasets to determine the correlation between cancer and pollutants for a given cluster/region in the USA. A) End-to-end workflow to determine the correlation between pollution and cancer data for certain clusters between 2018 and 2020. B) Steps to convert raw environmental data into a clean environmental dataset. C) Steps to convert raw cancer data into the cancer dataset.

https://doi.org/10.59720/24-322

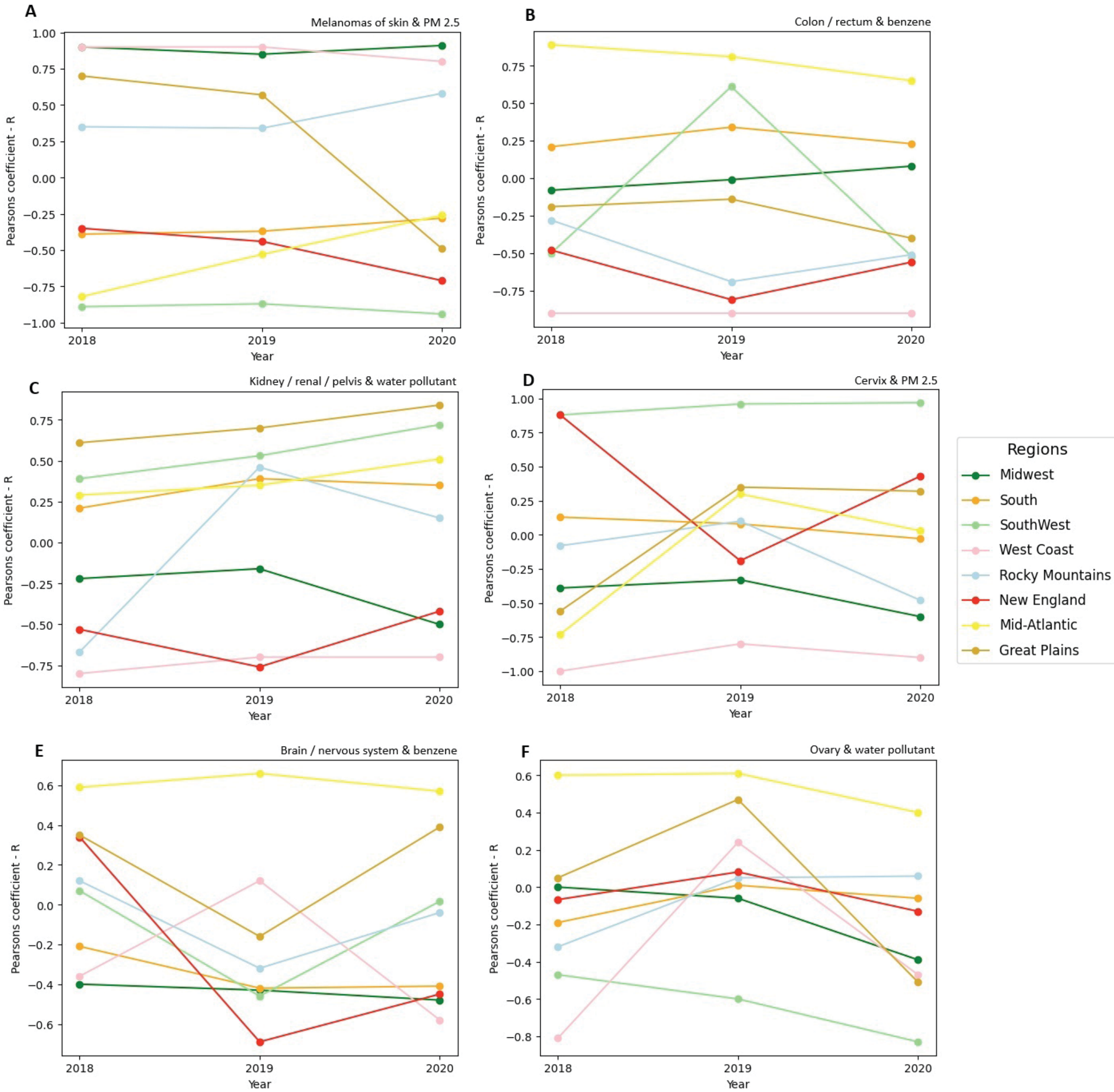


Figure 3: Graph representing correlation between cancer and regions for different pollutant types for the years 2018-2020. Strong correlations seen between certain cancer types and pollutants in specific regional clusters. A) Cancer type Melanomas of Skin and air pollutant PM2.5. B) Cancer type Colon and Rectum and chemical pollutant Benzene. C) Cancer type Kidney and Renal Pelvis and the water pollutant value of Total Pollutant Pounds (lb/yr) for Majors D) Cancer type Cervix and air pollutant PM2.5 E) Cancer type Brain and Other Nervous System and chemical pollutant Benzene. F) Cancer type Ovary and the water pollutant value of Total Pollutant Pounds (lb/yr) for Majors

Additionally, for the cancer dataset chosen, there were no publicly available data on certain types of cancer such as mouth cancer and eye cancer, which limited the scope of the study.

Additional methodological limitations should also be considered when interpreting these findings. Because this study relied on Pearson correlation analysis, the results identify statistical associations rather than causal relationships. Several potential confounding variables, including smoking prevalence, age distribution, socioeconomic status, healthcare access, occupational exposures, and genetic predispositions, may influence cancer incidence and could partially explain some of the observed correlations. Furthermore, the analysis treated each cancer type as a single category, despite many

https://doi.org/10.59720/24-322

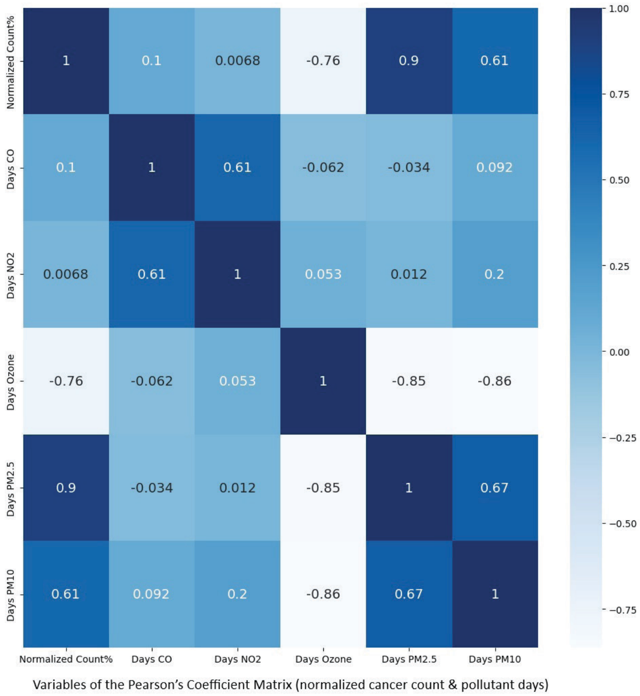


Figure 4: Example of Pearson coefficient matrix. The heatmap displays correlation values ranging from -1 to +1, where +1 (darkest blue) indicates the strongest positive correlation. Normalized cancer count is calculated as total cancer Incidence count in a region divided by its total population x 10000. This correlation matrix of environmental and health variables indicates a strong positive relationship ($r > 0.8$) between normalized skin melanoma cases and annual PM2.5 exposure days.

cancers consisting of biologically distinct subtypes that may have different environmental risk factors. For example, subtypes of lung, breast, and colorectal cancer may differ in their susceptibility to specific pollutants. Future studies could address these limitations by incorporating multivariable regression models or other advanced statistical approaches that account for potential confounders while examining individual pollutant–cancer relationships. Such approaches would provide a more comprehensive understanding of the extent to which environmental pollutants independently contribute to cancer incidence.

Environmental epidemiology holds significant potential for innovation and expansion. Validating our hypothesis by examining correlations in other countries could be a valuable next step. Further, the project scope could be broadened by incorporating additional diseases, cancer types, and environmental metrics. Conversely, narrowing the focus to a single pollutant and cancer type would enable a more comprehensive analysis, incorporating factors like race, gender, and income to better explain the Pearson correlation coefficient. As a next step, conducting the study at a city level may change accuracy by concentrating environmental factors and cancer types in a specific area.

Overall, the presence and strength of these correlations suggest that regional differences in environmental pollution are meaningfully associated with variation in cancer incidence across the country. Importantly, these patterns were consistent across the years analyzed, indicating that the observed relationships were not driven by short-term fluctuations. Collectively, these findings support the hypothesis that environmental pollutant exposure is associated with regional variation in cancer incidence and highlight the potential role of environmental factors in shaping public health outcomes.

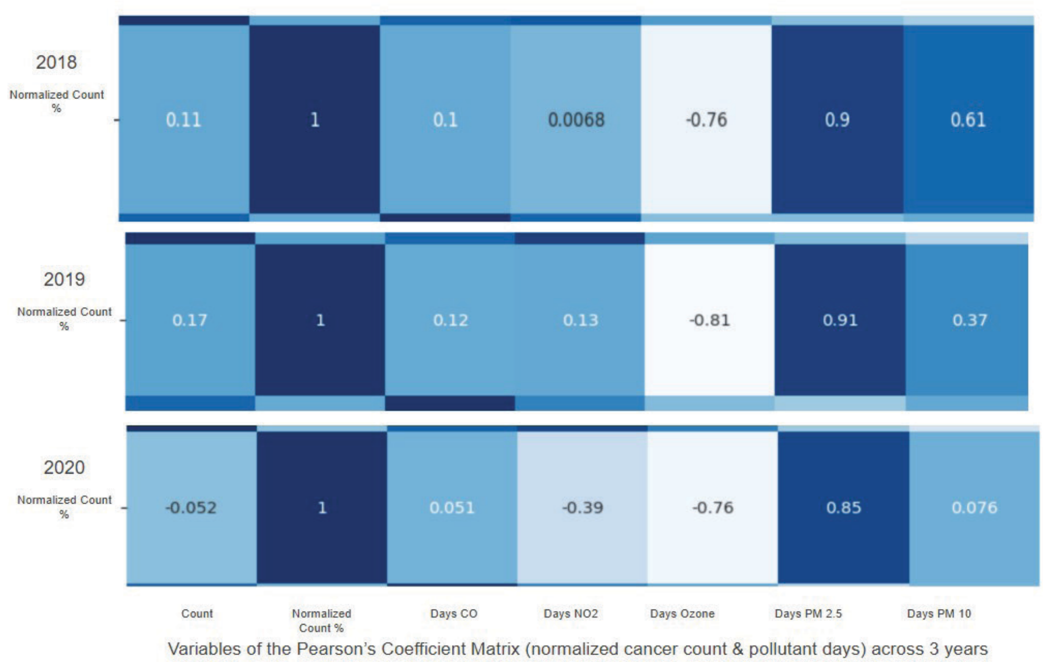


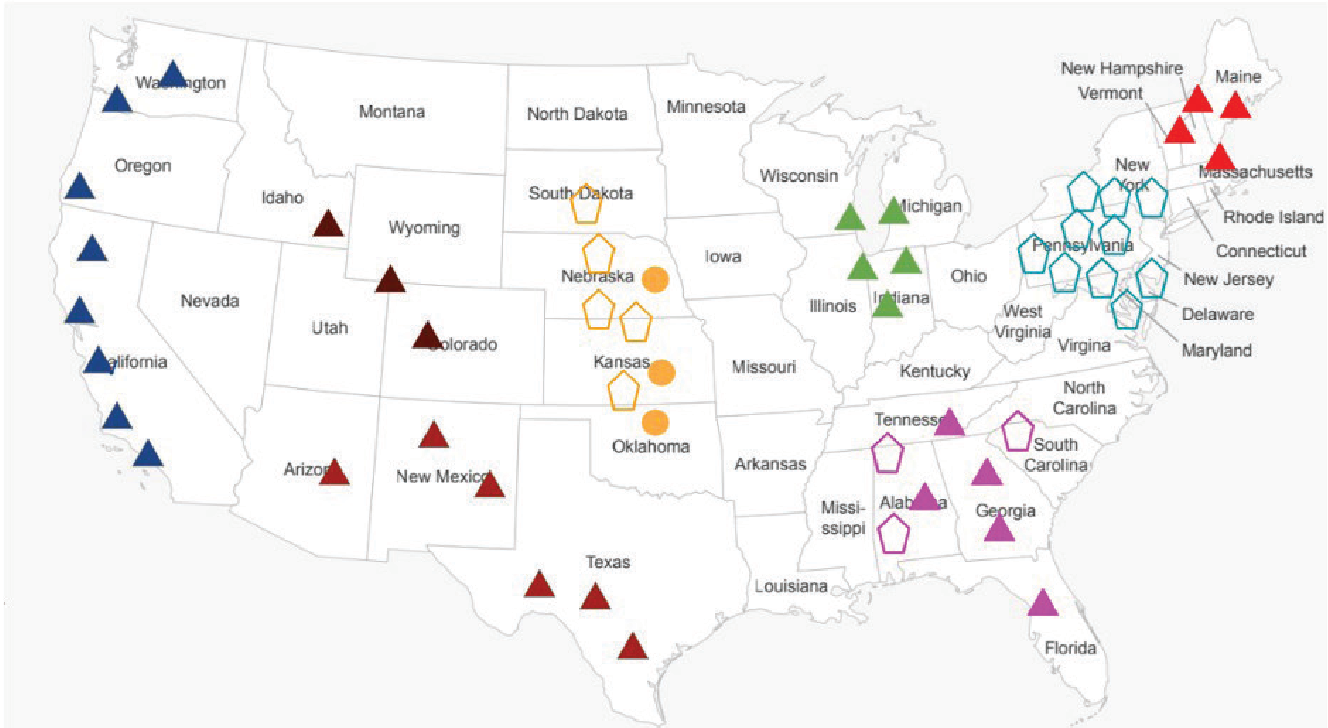
Figure 5: Evidence of strong cancer correlations between skin melanomas and air pollutant PM 2.5 in the Midwest. The correlation matrix of environmental and health variables from 2018–2020 indicates a strong positive relationship ($r > 0.8$) between normalized skin melanoma cases and annual PM 2.5 exposure days.

https://doi.org/10.59720/24-322

Region	States	Types of cancer	Cause	Pollutant type	Avg Cor-relation (2018-2020)	Test Statistic ($p < 0.05$)
Great Plains	Kansas, Nebraska, North Dakota, Oklahoma, South Dakota	Kidney and renal pelvis	Chemical	Benzene	0.867	0.0264
		Kidney and renal pelvis	Water	Total Pollutant Pounds (lb/yr) for Majors	0.717	0.13
Midwest	Illinois, Indiana, Iowa, Michigan, Minnesota, Missouri, Ohio, Wisconsin	Melanomas of the skin	Air	PM2.5	0.887	0.0098
Mid-Atlantic	Delaware, District of Columbia, Maryland, New Jersey, New York, Pennsylvania	Colon and rectum	Chemical	Benzene	0.783	0.0280
		Larynx	Chemical	Benzene	0.743	0.0453
New England	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont	Myeloma	Air	CO	0.8	0.0280
Rocky Mountains	Colorado, Idaho, Montana, Utah, Wyoming	Leukemias	Air	PM2.5	0.863	0.0279
West Coast	California, Oregon, Washington	Melanomas of the skin	Air	PM2.5	0.953	0.098
Southwest	Arizona, Nevada, New Mexico, Texas	Esophagus	Air	PM10	0.913	0.0427
		Cervix	Air	PM2.5	0.937	0.0315
South	Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, West Virginia	Hodgkin lymphoma	Air	Ozone	0.6	0.0196

Table 2: Strong correlation between various air pollutants and cancer across different geographical regions of the United States. Table showing strength of correlation and associated test statistics in different regions. A Python script was pre-processed to determine correlations in regions. (Determined by code - 46)

https://doi.org/10.59720/24-322



Legend

Symbol	Region	Types of Cancer	Cause	Pollutant Type
⬠	Mid-Atlantic	Colon and Rectum	Chemical	Benzene
⬠	Great Plains	Kidney and Renal Pelvis	Chemical	Benzene
●	Great Plains	Kidney and Renal Pelvis	Water	Total Pollutant Pounds (lb/yr) for Majors
▲	Midwest	Melanomas of the Skin	Air	PM2.5
▲	SouthWest	Esophagus	Air	PM10
▲	Rocky Mountains	Leukemias	Air	PM2.5
⬠	South	Stomach	Chemical	Chromium Compounds
▲	South	Hodgkin Lymphoma	Air	Ozone
▲	New England	Myeloma	Air	CO
▲	West Coast	Melanomas of the Skin	Air	PM2.5

▲ Air Pollutant
⬠ Chemical Pollutant
● Water Pollutant

*Different colors indicate different regions

Figure 6: Strong cancer correlations with pollutants across the different regions. Map showing distribution of the number of environmental factors that had a strong correlation to various cancers in that region. A Python script was pre-processed to determine correlations in regions. As listed in the legend, the various symbols highlight the type of pollutant causing cancer, and the different colors represent the different regions where that pollutant has a correlation with a cancer type. For example, the light-blue test represents the Mid-Atlantic region with blue pentagons representing a chemical pollutant that has caused a specific cancer cluster.

MATERIALS AND METHODS

Dataset Selection and Processing

To determine a correlation between pollutant metrics and cancer incidence rates, we first needed to obtain a reliable dataset for cancer incidence rates and these pollutants. We used Google Scholar, Perplexity.AI, and Google search to find raw cancer and environmental data. From this, we chose the EPA dataset for generating our environmental datasets and the CDC cancer dataset (18 – 21). We chose an EPA dataset for environmental data and a CDC dataset for cancer data because these sources had the most comprehensive data across many years at a state granularity. The EPA dataset contains raw data for air pollution, water pollution, and chemical pollution from 1980 to 2023 and considers over

https://doi.org/10.59720/24-322

200 different air pollution metrics, including NO_2, CO_2, ozone, PM10, lead, toxic waste in oceans at a state granularity (18, 19, 20). CDC has cancer datasets from 1980 to 2020 and counts the total number of cancer incidences as well as the population at a state granularity (2). Specifically, from the CDC cancer dataset, we used the variables YEAR, SITE, COUNT, EVENT_TYPE, and POPULATION to quantify annual cancer incidence by cancer type and to normalize incidence rates by population at the state level. From the EPA air pollution dataset, we selected state-level metrics representing the number of days per year with elevated concentrations of major air pollutants, including ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), particulate matter smaller than 2.5 µm (PM2.5), and particulate matter smaller than 10 µm (PM10). These pollutants were chosen because they are among the primary criteria air pollutants monitored nationwide by the EPA and have been previously associated with adverse health outcomes (**Table 1**). From the EPA chemical pollution dataset, we extracted variables corresponding to state identifiers (ST), total chemical releases (TOTAL RELEASES), and chemical type (CHEMICAL) to quantify the overall burden of industrial chemical emissions by state. In addition, from the EPA water pollution dataset, we used state-level measures of total pollutant discharge for major and non-major sources (lb/year) to capture variation in waterborne pollutant loads across regions. Together, these environmental variables provided a comprehensive representation of air, water, and chemical pollutant exposure used in the subsequent correlation analyses (**Table 1**). To pre-process (enhanced data quality of the raw data obtained from EDA and CDC), clean (removed unnecessary, erroneous values), and analyze the data, we used a custom-developed Python script.

Correlation Analysis

After normalizing and merging the cancer and environmental data, we used the Pearson correlation coefficient to determine the strength of the correlation between the prevalence of cancer incidences and air pollutants. A Pearson correlation coefficient matrix showcasing correlation between the different environmental pollutant variables and normalized cancer count for the Midwest region in 2018 was computed using Python code (46), where the different variables were compared to understand the correlation between them using the.corr() method of Pandas DataFrame and sns.heatmap(...) method in the Seaborn library (31, **Figure 4, Table 1**).

Data Normalization and Regional Aggregation

To normalize the cancer incidence rate, we created an additional column to account for the normalized count, calculated as count divided by population x 10000. Finally, we examined the correlation between environmental factors and disease-related indicators after performing the analysis for three years (**Figure 2**). The pandas, csv, seaborn, and matplotlib Python libraries were used with Kaggle as the platform to develop code (46) and plot all the graphs for this research analysis (31-34). 36 cancer types, across 8 regions in the USA, with 5 air pollutants, 2 water pollutants, and 5 chemical pollutants (12 total) were analyzed (**Table 1, 35**). A total number of 36 (number of cancer types) x 8 (number of regions) x 12 (number of pollutants) = 3456 correlation coefficient matrices were created and roughly 1.5 million rows of raw data processed into valid data frames. The flowchart describes the data procedures and data cleaning in more detail (**Figure 2A**). Extensive data cleaning and processing was employed to account for discrepancies in raw data when joining cancer and environmental datasets. This included removing unnecessary rows and columns, updating certain column data names, and removing unnecessary values to prepare the cancer dataset for correlation study (**Figure 2C**). In the pollution data frame, we selected the necessary columns (air pollutants/metrics) and grouped the data by each state (**Figure 2B**). Then, we created the cancer dataset. We selected the following columns: cancer count–new cancer cases in a year, population, and area (for a cancer type). To normalize the cancer incidence rate, we created an additional column to account for the normalized count, calculated as count divided by population x 10000. After obtaining these new data frames, we clustered states into eight geographic clusters, the Great Plains, New England, Southwest, Rocky Mountains, West Coast, Midwest, East, and Mid-Atlantic (**Figure 1**). Then, we mapped the environmental and cancer data into the different clusters. Then, we determined the strength of the correlation between the prevalence of cancer incidence rates and pollutants across these eight clusters using the Pearson correlation coefficient.

Statistical Significance Testing

To test our hypothesis, we used the Pearson Correlation Coefficient (outputs r value) as a mathematical tool to understand the strength of the correlation. r ϵ [-1, 1], with r ϵ (0.7, 1] being considered "strong. r ϵ [0.5 0.7] being considered "moderate". Test statistics were calculated corresponding to the r values and the region and if the associated *p*-value was less than 0.05, the correlation was considered statistically significant. Calculating the Pearson correlation coefficient alone does not confirm whether the observed correlation is statistically meaningful. Therefore, to ensure the validity of our findings, we computed a *t*-test for correlation to determine if the observed r values were likely due to chance variation in our sample sizes. By converting r to *t*-statistics and using the *t*-distribution to compute a *p*-value, we assessed the statistical significance of each relationship. This approach strengthened our analysis by distinguishing strong, reliable correlations from those that may occur randomly, which is especially important given the relatively small sample sizes within each region. For statistical significance, we calculated a test statistic for each correlation using the equation:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

where *r* is the correlation coefficient and *n* is the sample size (number of states in a region). The resulting *t*-value was used to compute a *p*-value using the *t*-distribution with (n - 2) degrees of freedom. Correlations with $p < 0.05$ were considered statistically significant.

Received: September 1, 2024
Accepted: August 20, 2025
Published: September 12, 2026

https://doi.org/10.59720/24-322

REFERENCES

1. Siegel, Rebecca L., *et al.* "Cancer Statistics, 2022." *CA: A Cancer Journal for Clinicians*, vol. 72, no. 1, 2022, pp. 7–33. https://doi.org/10.3322/caac.21708.
2. "Cancer Data and Statistics." Centers for Disease Control and Prevention. www.cdc.gov/cancer/data/index.html. Accessed 18 Aug. 2024.
3. Cronin, Kathleen A., *et al.* "Annual Report to the Nation on the Status of Cancer, Part 1: National Cancer Statistics." *Cancer*, vol. 128, no. 24, 15 Dec. 2022, pp. 4251–4284. https://doi.org/10.1002/cncr.34479.
4. "Risk Factors: Age." National Cancer Institute. www.cancer.gov/about-cancer/causes-prevention/risk/age. Accessed 18 Aug. 2024.
5. "Water Pollution." *National Geographic Society*. https://education.nationalgeographic.org/resource/pollution/. Accessed 17 June 2025.
6. Pope III, C. Arden, *et al.* "Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution." *JAMA*, vol. 287, no. 9, 6 Mar. 2002, pp. 1132–1141. https://doi.org/10.1001/jama.287.9.1132.
7. Ward, Mary H., *et al.* "Drinking Water Nitrate and Human Health: An Updated Review." *International Journal of Environmental Research and Public Health*, vol. 15, no. 7, 23 July 2018, p. 1557. https://doi.org/10.3390/ijerph15071557.
8. Spatari, Giovanna, *et al.* "Epigenetic Effects of Benzene in Hematologic Neoplasms: The Altered Gene Expression." *Cancers*, vol. 13, no. 10, 2021, article 2392. https://doi.org/10.3390/cancers13102392.
9. "New Report: Nearly Half of People in U.S. Exposed to Dangerous Air Pollution Levels." American Lung Association. www.lung.org/media/press-releases/state-of-the-air-2025. Accessed 16 June 2026.
10. "Greenhouse Gases Continued to Increase Rapidly in 2022." NOAA Climate.gov. www.climate.gov/news-features/feed/greenhouse-gases-continued-increase-rapidly-2022. Accessed 16 June 2026.
11. Scheringer, Martin, and Ralf Schulz. "The State of the World's Chemical Pollution." *Annual Review of Environment and Resources*, vol. 50, 2025, pp. 381–408. https://doi.org/10.1146/annurev-environ-111523-102318.
12. Levin, Ronnie, *et al.* "US Drinking Water Quality: Exposure Risk Profiles for Seven Legacy and Emerging Contaminants." *Journal of Exposure Science & Environmental Epidemiology*, vol. 34, no. 1, 2024, pp. 3–22. https://doi.org/10.1038/s41370-023-00597-z.
13. "Lung Cancer Key Findings." American Lung Association. www.lung.org/research/state-of-lung-cancer/key-findings. Accessed 16 June 2026.
14. Christian, W. Jay, *et al.* "Exploring Geographic Variation in Lung Cancer Incidence in Kentucky Using a Spatial Scan Statistic: Elevated Risk in the Appalachian Coal-Mining Region." *Public Health Reports*, vol. 126, no. 6, 2011, pp. 789–796. https://doi.org/10.1177/003335491112600604.
15. "State Cancer Profiles > Incidence Rates Table." *State Cancer Profiles*. National Cancer Institute and Centers for Disease Control and Prevention. www.statecancerprofiles.cancer.gov/incidencerates/index.php?age=001&areatype=state&cancer=020&race=00&sex=0&sortOrder=desc&sortVariableName=rate&stage=999&stateFIPS=00&type=incd&year=1. Accessed 16 June 2026.
16. Kolpin, D. W., *et al.* "Nitrate in Groundwater of the Midwestern United States: A Regional Investigation on Relations to Land Use and Soil Properties." *IAHS-AISH Publication*, no. 257, 1999, pp. 111–116. https://pubs.usgs.gov/publication/70020921. Accessed 16 June 2026.
17. "Water Contaminants and Cancer Risk: Arsenic, Disinfection Byproducts, and Nitrate." National Cancer Institute. www.dceg.cancer.gov/research/what-we-study/drinking-water-contaminants. Accessed 16 June 2026.
18. Groves, F. D., *et al.* "Is There a 'Cancer Corridor' in Louisiana?" *Journal of the Louisiana State Medical Society*, vol. 148, no. 4, Apr. 1996, pp. 155–165.
19. Robinson, Ellis S., *et al.* "Ethylene Oxide in Southeastern Louisiana's Petrochemical Corridor: High Spatial Resolution Mobile Monitoring during HAP-MAP." *Environmental Science & Technology*, vol. 58, no. 25, 2024, pp. 11084–11095. https://doi.org/10.1021/acs.est.3c10579.
20. "U.S. Cancer Statistics Data Visualizations Tool: Data Tables." Centers for Disease Control and Prevention. www.cdc.gov/united-states-cancer-statistics/dataviz/data-tables.html. Accessed 16 June 2026.
21. "State Statistics." U.S. Environmental Protection Agency. https://echo.epa.gov/trends/loading-tool/get-data/state-statistics. Accessed 16 June 2026.
22. "TRI Basic Data Files: Calendar Years 1987–Present." U.S. Environmental Protection Agency. www.epa.gov/toxics-release-inventory-tri-program/tri-basic-data-files-calendar-years-1987-present. Accessed 16 June 2026.
23. "AirData Download Files." U.S. Environmental Protection Agency. https://aqs.epa.gov/aqsweb/airdata/download_files.html. Accessed 16 June 2026.
24. Python Software Foundation. *Python* Version 3.12.2, Python Software Foundation, 2024, www.python.org/.
25. Jones, Rena R., et al. "Nitrate from Drinking Water and Diet and Bladder Cancer Among Postmenopausal Women in Iowa." *Environmental Health Perspectives*, vol. 124, no. 11, Nov. 2016, pp. 1751–1758. https://doi.org/10.1289/EHP191.
26. Temkin, Alexis M., *et al.* "Exposure-Based Assessment and Economic Valuation of Adverse Birth Outcomes and Cancer Risk Due to Nitrate in United States Drinking Water." *Environmental Research*, vol. 176, Sept. 2019, article 108442. https://doi.org/10.1016/j.envres.2019.04.009.
27. Hughes, Michael F., *et al.* "Arsenic Exposure and Toxicology: A Historical Perspective." *Toxicological Sciences*, vol. 123, no. 2, Oct. 2011, pp. 305–332. https://doi.org/10.1093/toxsci/kfr184.
28. Vienne-Jumeau, A., *et al.* "Environmental Risk Factors of Primary Brain Tumors: A Review." *Revue Neurologique*, vol. 175, no. 10, 2019, pp. 664–678. https://doi.org/10.1016/j.neurol.2019.08.004.
29. "Determining If Something Is a Carcinogen." American Cancer Society. www.cancer.org/cancer/risk-prevention/understanding-cancer-risk/determining-if-something-is-a-carcinogen.html. Accessed 16 June 2026.
30. Maisara, Siti, *et al.* "Air Pollution and Its Association with

https://doi.org/10.59720/24-322

Cervical Cancer: A Scoping Review." *Medicine & Health*, vol. 18, no. 1, 2023, pp. 9–20. https://doi.org/10.17576/mh.2023.1801.03.
31. Waskom, Michael L. "seaborn: Statistical Data Visualization." *Journal of Open Source Software*, vol. 6, no. 60, 2021, p. 3021. https://doi.org/10.21105/joss.03021.
32. The pandas development team. *pandas* Version 2.2.2. NumFOCUS, 2024. https://pandas.pydata.org/.
33. Harris, Charles R., *et al.* "Array Programming with NumPy." *Nature*, vol. 585, no. 7825, 2020, pp. 357–362. https://doi.org/10.1038/s41586-020-2649-2.
34. Hunter, J. D. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering*, vol. 9, no. 3, 2007, pp. 90–95. https://doi.org/10.1109/MCSE.2007.55.
35. Alexander, Kathy. "Geographic Summary of the United States." *Legends of America*. www.legendsofamerica.com/ah-geosum/. Accessed 18 Aug. 2024.
36. "Carbon Monoxide and Health." California Air Resources Board. www.arb.ca.gov/resources/carbon-monoxide-and-health. Accessed 17 June 2025.
37. "Nitrogen Dioxide." American Lung Association. www.lung.org/clean-air/outdoors/what-makes-air-unhealthy/nitrogen-dioxide. Accessed 17 June 2025.
38. "Health Effects of Ozone Pollution." U.S. Environmental Protection Agency. www.epa.gov/ground-level-ozone-pollution/health-effects-ozone-pollution. Accessed 17 June 2025.
39. "Inhalable Particulate Matter and Health." California Air Resources Board. https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health. Accessed 17 June 2025.
40. "Discharge Monitoring Report (DMR) Search Results Help." U.S. Environmental Protection Agency. https://echo.epa.gov/help/loading-tool/water-pollution-search/search-results-help-dmr. Accessed 17 June 2025.
41. "Health Effects of Lead." U.S. Environmental Protection Agency. www.epa.gov/lead/what-are-some-health-effects-lead. Accessed 17 June 2025.
42. "Sulfuric Acid." New Jersey Department of Health. https://nj.gov/health/eoh/rtkweb/documents/fs/1761.pdf. Accessed 17 June 2025.
43. "Hexavalent Chromium." National Cancer Institute. www.cancer.gov/about-cancer/causes-prevention/risk/substances/chromium. Accessed 17 June 2025.
44. "Benzene and Cancer Risk." *American Cancer Society*. www.cancer.org/cancer/risk-prevention/chemicals/benzene.html. Accessed 17 June 2025.
45. "Polycyclic Aromatic Hydrocarbons (PAHs)." Agency for Toxic Substances and Disease Registry (ATSDR). Centers for Disease Control and Prevention. https://wwwn.cdc.gov/TSP/PHS/PHS.aspx?phsid=120&toxid=25. Accessed 17 June 2025.
46. Link to code on GitHub (hosted on Kaggle): NoblePenguin43. (n.d.). *Cancer-environment correlation* [Computer software]. GitHub. https://github.com/noblepenguin43/cancer-environment-corelation