Environ Geochem Health (2024) 46:82
https://doi.org/10.1007/s10653-023-01820-4

ORIGINAL PAPER

Evaluating county-level lung cancer incidence from environmental radiation exposure, $PM_{2.5}$, and other exposures with regression and machine learning models

Heechan Lee · Heidi A. Hanson · Jeremy Logan · Dakotah Maguire · Anuj Kapadia · Shaheen Dewji · Greeshma Agasthya

Received: 18 July 2023 / Accepted: 27 November 2023 / Published online: 17 February 2024

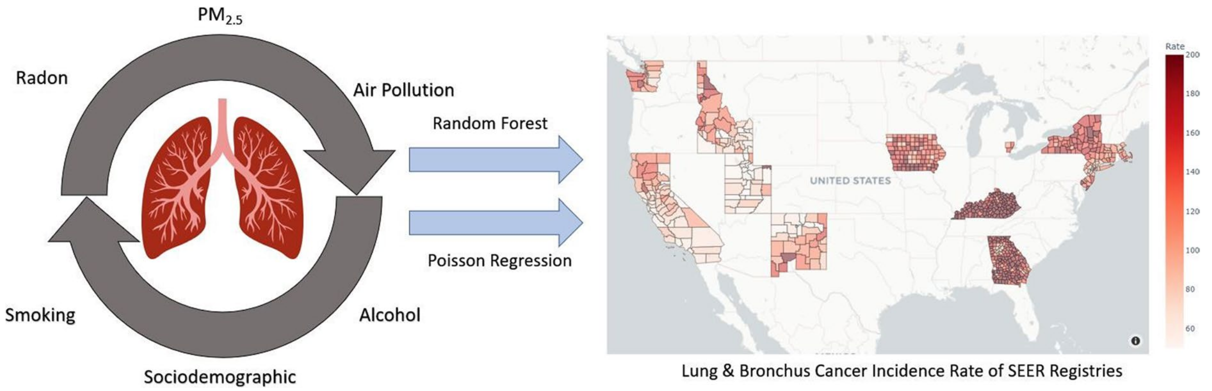
Abstract Characterizing the interplay between exposures shaping the human exposome is vital for uncovering the etiology of complex diseases. For example, cancer risk is modified by a range of multifactorial external environmental exposures. Environmental, socioeconomic, and lifestyle factors all shape lung cancer risk. However, epidemiological studies of radon aimed at identifying populations at high risk for lung cancer often fail to consider multiple exposures simultaneously. For example, moderating factors, such as $PM_{2.5}$, may affect the transport of radon progeny to lung tissue. This ecological analysis leveraged a population-level dataset from the National Cancer Institute's Surveillance, Epidemiology, and End-Results data (2013–17) to simultaneously investigate the effect of multiple sources of low-dose radiation (gross γ activity and indoor radon) and $PM_{2.5}$ on lung cancer incidence rates in the USA. County-level factors (environmental, sociodemographic, lifestyle) were controlled for, and Poisson regression and random forest models were used to assess the association between radon exposure and lung and bronchus cancer incidence rates. Tree-based machine learning (ML) method perform better than traditional regression: Poisson regression: 6.29/7.13 (mean absolute percentage error, MAPE), 12.70/12.77 (root mean square error, RMSE); Poisson random forest regression: 1.22/1.16 (MAPE), 8.01/8.15 (RMSE). The effect of $PM_{2.5}$ increased with the concentration of environmental radon, thereby confirming findings from previous studies that investigated the possible synergistic effect of radon and $PM_{2.5}$ on health outcomes. In summary, the results demonstrated (1) a need to consider multiple environmental exposures when assessing radon exposure's association with lung cancer risk, thereby highlighting (1) the importance of an exposomics framework and (2) that employing ML models may capture the complex interplay between environmental exposures and health, as in the case of indoor radon exposure and lung cancer incidence.

H. Lee · S. Dewji
Nuclear and Radiological Engineering and Medical Physics Programs, George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, 770 State Street, Atlanta, GA 30332, USA

H. Lee · H. A. Hanson · D. Maguire · A. Kapadia · G. Agasthya (✉)
Advanced Computing for Health Sciences Section, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37830, USA
e-mail: hansonha@ornl.gov

J. Logan
Data Engineering Group, Data and AI Section, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37830, USA

Graphical abstract



Keywords Ionizing radiation · Lung cancer · Radon · Exposome · $PM_{2.5}$

Introduction

The concept of the human exposome was first proposed almost two decades ago as a framework to guide research that explores the etiological complexities of health and disease (Wild, 2005). Because the relationship between multifactorial exposure patterns that influence health outcomes is complex, there is a need for studies that incorporate information from multiple exposures. Approaches that include a single environmental exposure may not fully or accurately describe the risk of disease because mixing factors may alter the effects of a single exposure (Wild, 2005; Zhang et al., 2021). Alpha radiation, consisting of two protons and two neutrons, can be easily stopped by skin or paper, yet is harmful if ingested. Beta radiation, comprising electrons or positrons, can also be readily halted but poses risks to the human body when ingested. Gamma radiation, a high-energy electromagnetic wave, is produced by nuclear reactions and has strong penetrative capabilities; therefore, external exposure can cause significant harm. For example, recent studies have shown that known associations between fine particulate matter ($PM_{2.5}$) and health are modified by gross β activity (a measure of the counts of β ray per unit time) (Blomberg et al., 2019; Dong et al., 2022). In the present study, we leverage statistical and ML methods to simultaneously consider the effects of background radiation levels (gamma radiation emitted from airborne radioactive particles and indoor radon gas), $PM_{2.5}$ exposure, and other social and behavioral factors on county-level lung cancer rates.

The effects of radiation on health have long been investigated. Since the discovery of radiation and its subsequent applications in the nuclear fuel cycle, industry, security and consequence management, and nuclear medicine, the study of radiation-induced health effects on humans has been an important field of research. The evaluation of the risk from radiation exposure at high doses/high dose rates has been relatively well established through the Life Span Study of atomic bomb survivors (Seong et al., 2016; Ozasa et al., 2019; United Nations Scientific Committee on the Effects of Atomic Radiation [UNSCEAR], 2008), the Chernobyl workers study (Morton et al., 2021), and nuclear weapon fallout studies (Lyon et al., 2006; Takahashi et al., 2003). However, in determining radiation-induced outcomes, less is known about low-level or background exposures and their interactions with other environmental toxins.

Radon Exposure and Lung Cancer

Radon is produced from the decay of uranium and is emitted from terrestrial sources and building materials. It is known that exposure to radon is affected by various factors such as housing characteristics, surficial uranium concentration, soil permeability, and even groundwater (Mose & Mushrush, 1999; Ponciano-Rodríguez et al., 2021; Przylibski et al., 2022; Smith & Field, 2007). This alpha emitter and its progenies are absorbed into the lungs, which in the process exposes the airways to radiation. Radionuclides inhaled into the lungs and bronchi cause the ionization of biological molecules, which in turn causes DNA damage and potentially cancer (Abergel et al., 2022; National Research Council [NRC], 1999; McDonald et al., 1995). In this process, most of the inhaled radon is exhaled from the lung because radon is an inert gas. However, its progenies have short half-lives that usually decay before they are removed from the lung through exhalation (Tirmarche et al., 2010). Additional complexities arise when considering the other particles and gases that the radon progeny bind to, such as $PM_{2.5}$. The pollution particles may serve as a vector for the deposition of radon progeny into the lungs.

Epidemiological studies that link radon exposure to lung cancer risk show conflicting results (Cohen & Colditz, 1994; Kreuzer et al., 2010, 2015; Mifune et al., 1992; Yoon et al., 2016). Although there is strong evidence from studies of uranium miners (Richardson et al., 2021), studies of the broader population are less conclusive. A recent ecological study conducted in Mexico examined the relationship between indoor radon exposure and lung cancer mortality. The findings of the study suggested that higher levels of radon concentration may be linked to an increased risk of lung cancer (Ponciano-Rodríguez et al., 2021). However, this study has a limitation in that it did not control variables such as lifestyle or socioeconomic status in the model. In another study based in the USA, Cohen et al. (1994, 1995) used county-level data to investigate the association of radon exposure and lung cancer. This study showed a negative association between lung cancer risk and radon concentration, even though Cohen controlled for several confounding factors, such as smoking, socioeconomic factors, and geography (Cohen, 1995; Cohen & Colditz, 1994). There are several limiting factors within the aforementioned studies, for example, small sample sizes and challenges associated with decoupling the risk associated with radon exposure and other confounding factors (e.g., lifestyle, socioeconomic factors).

New computational methods for population-level exposomic research

The emerging fields of data science and ML provide new opportunities for characterizing the relationship between the exposome and lung cancer by offering alternative analytical methods for modeling complex relationships between social and environmental determinants of health. Using ML methods for modeling complex relationships in epidemiology research has become increasingly prevalent (Wiemken & Kelley, 2020). This study tested the effects of low-dose radiation in single- and multi-exposure models and compared results from traditional methods and ML methods for a comprehensive look at the relationship between low-dose radiation and lung cancer rates in the USA.

Methodology

This ecological analysis utilized county-level data to describe the relationship between two measures of low-dose radiation exposure and lung cancer rates. The following county-level factors were assembled and are summarized in Table 1: (1) environmental radiation exposures (gross gamma activity and indoor radon), (2) non-radiation environmental variables (air quality), (3) lifestyle (smoking), and (4) sociological data (demographic/socioeconomic). These variables were used to predict county-level lung/bronchus cancer risk and incidence. We tested for multicollinearity and all of the variables showed variance inflation factor (VIF) less than 5. Poisson regression and Poisson random forest (RF) regression were used to model lung cancer incidence rates in 662 counties in the USA. The MAPE (Mean Absolute Percentage Error) and RMSE (Root Mean Square Error) from a fivefold cross-validation were compared across regression models to analyze model performance. The codes used in this analysis can be found in https://github.com/Heechan-Lee/county_radon_lung.cancer.

Environmental radiation data

Gamma count rate

The US Environmental Protection Agency's (EPA's) RadNet system monitors the gamma counts across

Table 1 Radiation, environmental, sociological, and cancer incidence datasets of 662 counties in the USA used in this study

Model grouping	Data	Source	Year(s)	Data description
RadNet	RadNet	EPA (EPA, n.d.)	2-year lagged averaged	Gamma count rate (interpolated & averaged), cpm
Radon	Radon zone	EPA (EPA, 1993)	1993	Zone 1: > 4 pci/L Zone 2: 2–4 pci/L Zone 3: < 2 pci/L
Radon	Radon concentration	CDC (CDC, n.d.)	2008–2017	Indoor radon tests from labs (median), pCi/L
$PM_{2.5}$	Air quality, $PM_{2.5}$	CDC (CDC, n.d.)	3-year lagged averaged	$PM_{2.5}$ (averaged), μg/m^3
Others	Air quality	CDC (CDC, n.d.)	2011	Chemicals (formaldehyde, benzene), μg/m^3
Others	Air quality, ozone	CDC (CDC, n.d.)	3-year lagged averaged	Ozone (averaged), days 8-h average ozone concentration exceeded 0.07 ppm
Smoking	Smoking	CHR (University of Wisconsin Population Health Institute, 2022)	2015	Adult smoking rate, %
Sociodemographic	Education	SEER (NCI-DCCPS-SRP, 2022)	2008–2012	% of the population with high school education
Sociodemographic	Income	SEER (NCI-DCCPS-SRP, 2022)	2008–2012	Median family income, USD
Sociodemographic	Unemployment	SEER (NCI-DCCPS-SRP, 2022)	2008–2012	Rate, %
Sociodemographic	Urban %	SEER (NCI-DCCPS-SRP, 2022)	2010	% of the total population in urban areas
Population	Population	SEER (NCI-DCCPS-SRP, 2022)	2013–2017	Total population
Health outcomes	Lung (bronchi) cancer incidence	SEER (NCI-DCCPS-SRP, 2022)	2013–2017	Incidence count per age group and sex

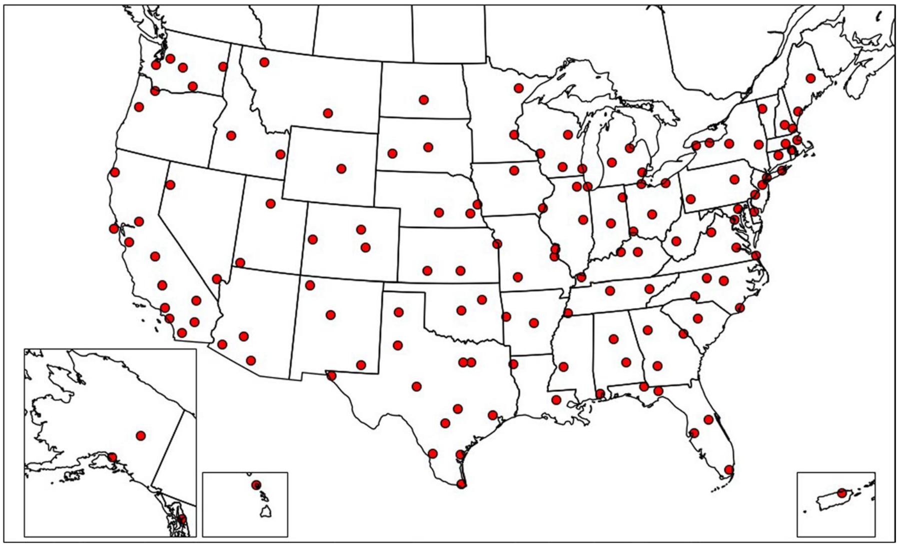


Fig. 1 Locations of 140 counties (or equivalent) that have RadNet monitoring centers in the USA and Puerto Rico

the United States (Fraass, 2015). The first monitoring center came online in 2006, and since then the number of monitoring centers has increased to 140. (Fig. 1) Since July 2016, 80 monitoring centers also record the gamma exposure rate; however, this data is not available for the entire timeframe of the cancer incidence data, so the gamma gross count rate is used instead. Gamma gross count rates are measurements of radiation emitted from a particulate collected on an air filter—they are not a direct measure of exposure rate.

The three most prominent limitations of this dataset were as follows: (1) a high percentage of the monitoring centers were missing data from one or more months, (2) the data were limited to 140 county data points, and (3) some of the monitoring centers did not have records prior to 2013. To overcome these limitations, data were imputed by using existing alternate datasets. First, the monthly average hourly-reported gamma gross count was calculated to capture the seasonality of the data and to minimize the effect of outliers caused by local volatility. Second, the following two imputations were implemented: (1) imputing data of missing months through linear interpolation and seadec (Seasonally Decomposed Missing Value Imputation) function of R (R Core Team, 2021) from the imputeTS package(Moritz & Bartz-Beielstein, 2017) and (2) 2D linear interpolation by using the 'griddata' function of SciPy from Python (Virtanen et al., 2020). Linear interpolation and the seadec function outperformed interpolation with mean value as well as other methods, such as ARIMA (Autoregressive Integrated Moving Average) with Kalman filters and seasplit (Seasonally Splitted Missing Value Imputation) function of imputeTS package (Moritz & Bartz-Beielstein, 2017), for imputing the missing months. For 2D interpolation of counties without nearby or inherent geographical obstacles such as mountains or large forests, the interpolation showed less than 15% percentage error between the averaged real and predicted counts. The map of interpolated gamma count rate data is shown in Fig. 2.

We then created a summary measure of gross gamma activity for each observation year by averaging the gross gamma activity for the two years prior to the year of diagnosis (two-year lag).

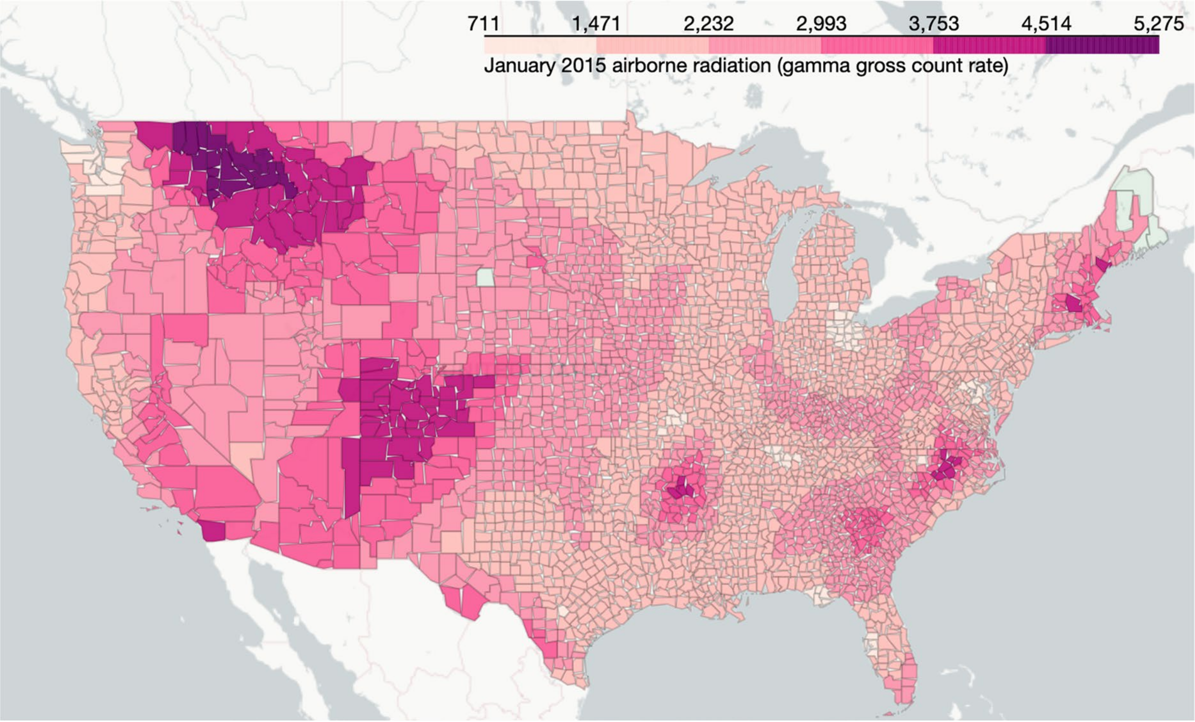


Fig. 2 Interpolated gamma count rate (RadNet) data for the USA from data of 140 monitoring centers created with seadec function and linear interpolation

Radon

Two sources of radon data were utilized in this study: radon zones from the US Environmental Protection Agency (1993) and the median concentration from indoor radon test kits downloaded from the US Centers for Disease Control and Prevention (CDC) database (Centers for Disease Control and Prevention, n.d.). The radon zone data classification was developed by the EPA in 1993 and classifies counties into three groups based on the potential for exposure to indoor radon: **Zone 1**, representing the highest radon concentration group of 4 pCi/L or higher; **Zone 2**, with a radon concentration of 2–4 pCi/L; and **Zone 3**, with less than 2 pCi/L (US Environmental Protection Agency, 1993). Although the classification system is almost three decades old, the classifications can reasonably be assumed as representative of the composition of soil and bedrock, which do not change significantly over this elapsed time. The CDC radon data are the results of indoor radon tests from kits deployed in residential, industrial, and educational locations across the USA from 2008 to 2017. The locations of 662 counties used in the analysis and the average of yearly median indoor radon concentrations of those counties are shown in Fig. 3.

Non-radiation environmental data

Air quality

Air pollution, notably particulate matter, is a known lung cancer-inducing factor (Couraud et al., 2012; Dela Cruz et al., 2011; Raaschou-Nielsen et al., 2013; Turner et al., 2011a, 2011b). Air quality-based measurements were obtained from the National Environmental Public Health Tracking Network (CDC, n.d.). This database includes various features, including toxic chemicals, ozone, and $PM_{2.5}$. For toxic chemicals, the measurements were of an annual average concentration of 2005 and 2011. Ozone data was based on the days that the daily 8-h average ozone concentration exceeded 0.07 ppm between 2001 and 2016. The $PM_{2.5}$ data was the average concentration of $PM_{2.5}$ for each year between 2001 and 2016. For data on chemical concentrations, data from 2011 was extracted to best align with cancer incidence data, and the concentrations were assumed not to have changed by 2017. Among the various chemicals, formaldehyde

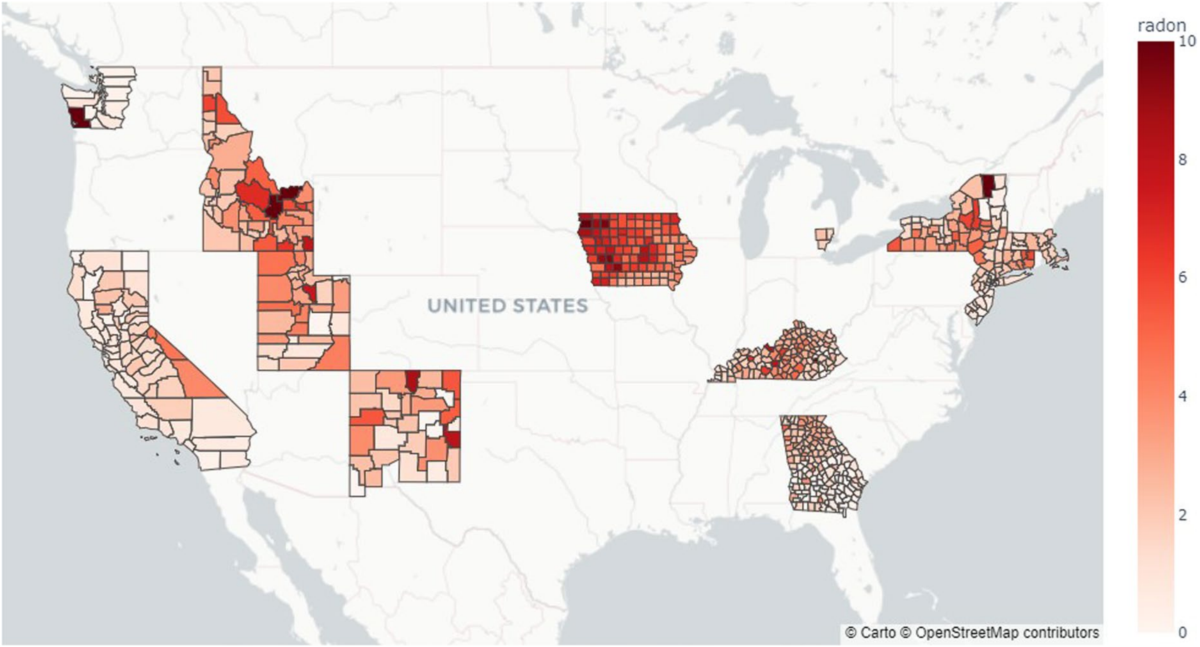


Fig. 3 Average of yearly median indoor radon concentration for US Counties with SEER data downloaded from the US Centers for Disease Control and Prevention in pCi/L

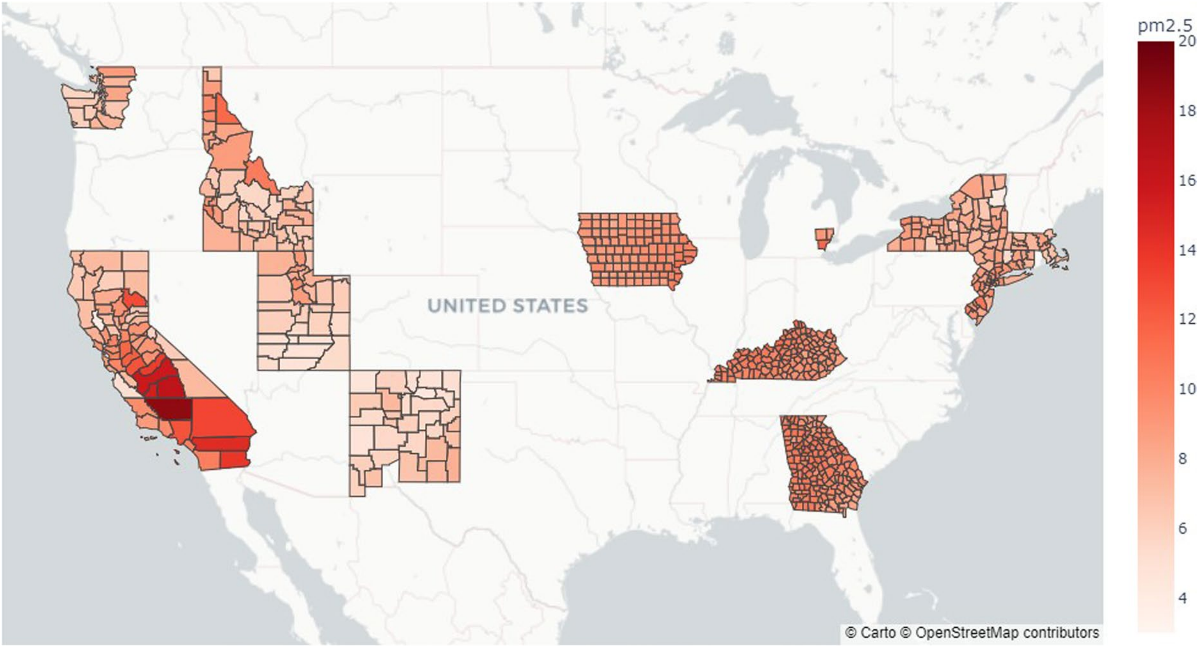


Fig. 4 Average of $PM_{2.5}$ concentration for US Counties with SEER cancer incidence data in μg/m^3

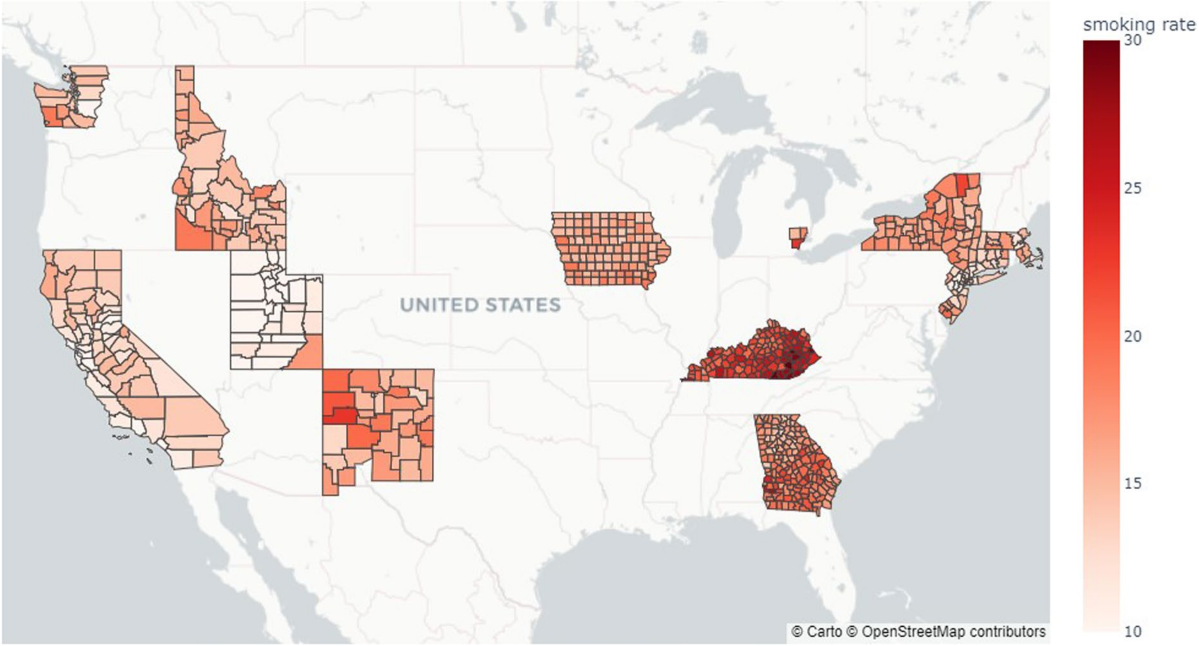


Fig. 5 The percentage of adults in US Counties with SEER cancer incidence data who self-identified as smokers in a 2015 state-based random digit dial telephone survey of the Behavioral Risk Factor Surveillance System

and benzene were employed. For ozone and $PM_{2.5}$, the average of the concentrations for the 3 years preceding the cancer incidence record was included in this analysis. The locations of 662 counties and their average $PM_{2.5}$ concentrations are shown in Fig. 4.

Lifestyle data

Tobacco smoking

Smoking has been well established as the leading lung cancer-inducing factor (de Groot et al., 2018; Dela Cruz et al., 2011). The smoking data included in this study was adapted from the Robert Wood Johnson Foundation County Health Rankings (CHR) (University of Wisconsin Population Health Institute [UWPHI], 2022; Remington et al., 2015). This dataset provides the percentage of adults who self-identified as smokers in a 2015 state-based random digit dial telephone survey of the Behavioral Risk Factor Surveillance System. The 2015 smoking rates were chosen because smoking rates are relatively constant during the period of observation, and the midpoint of interest was used as the representative year. The map of the smoking rates in 2015 are depicted in Fig. 5.

Demographic and socioeconomic data

Demographic and socioeconomic factors from the Surveillance, Epidemiology, and End Results (SEER) database (National Cancer Institute, DCCPS, Surveillance Research Program [NCI-DCCPS-SRP], 2022) were included in the dataset because cancer incidence is affected by various demographic and socioeconomic factors (Siegel et al., 2019). The SEER data includes education level, poverty rate, unemployment rate, rate of residence in urban areas, and divided into age cohorts separated by 5 years to reflect the age effect. Age groups range from 30–34 years old to 80–84 years old for both sexes. Age groups range from 30–34 years old to 80–84 years old for both sexes. The averages from 2008 to 2012 reported for each county were used for high school education, median family income, and unemployment data and were assumed to remain constant. The urban rate was taken from 2010 data. Total population from the 2010 US Census for the 662 countres used in the analysis are shown in Fig. 6.

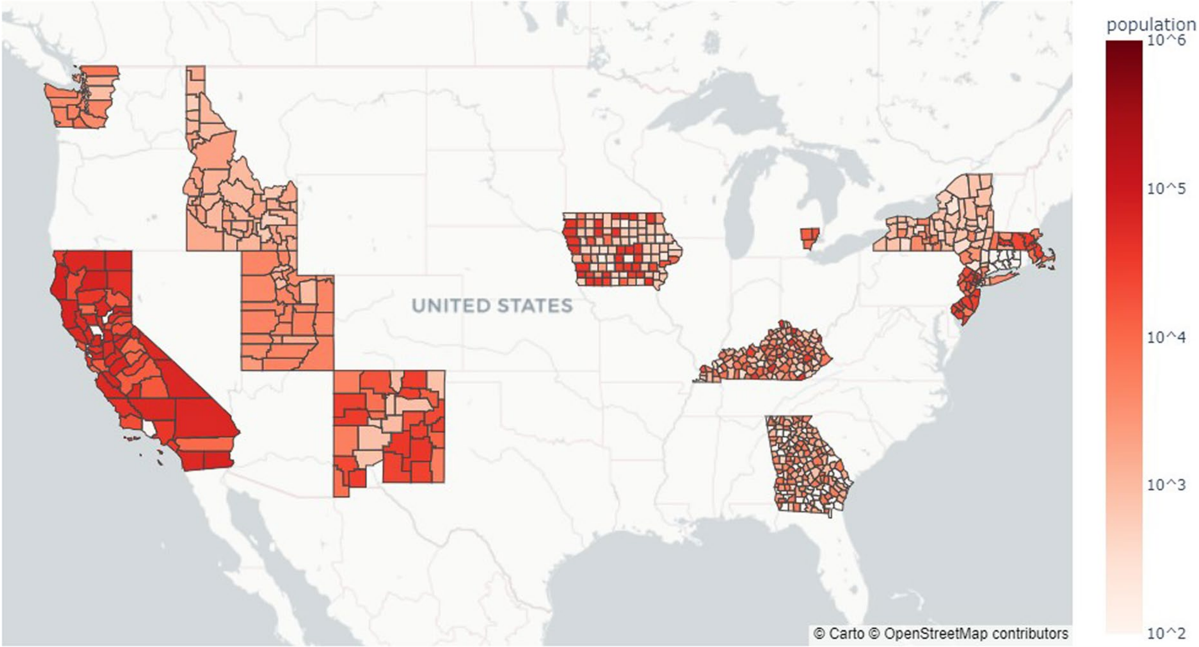


Fig. 6 Total population in 2010 for counties with SEER cancer incidence data

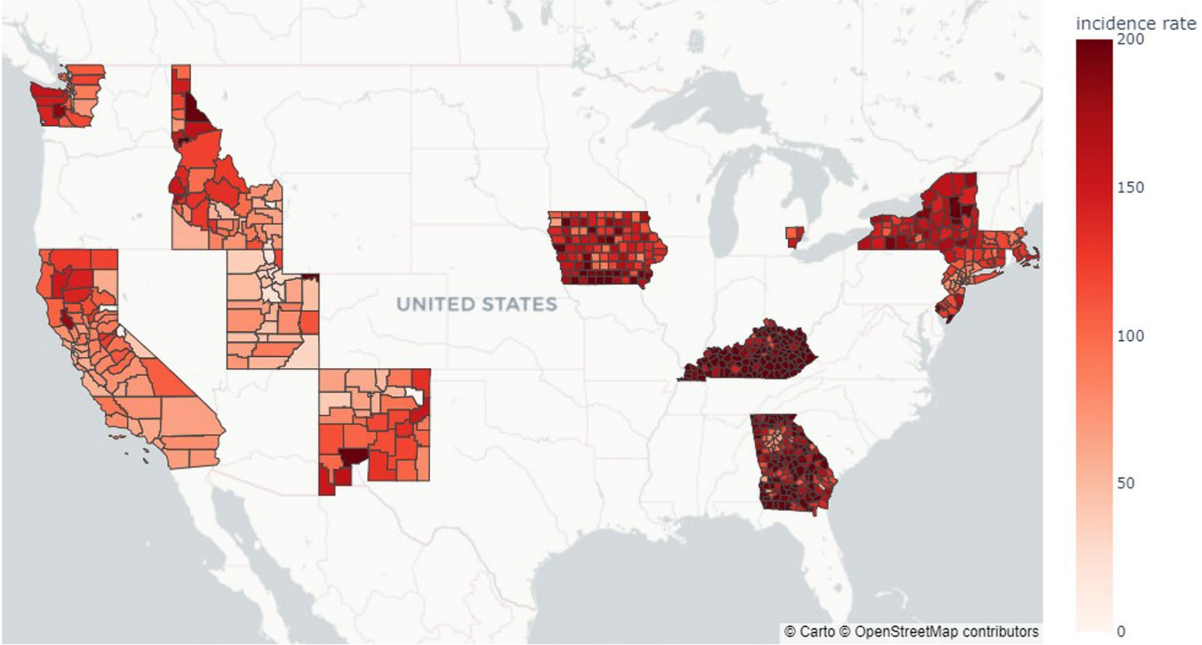


Fig. 7 Cancer incidence rate of age-groups of interests (incidence per 100,000)

Health outcomes

Lung and bronchus cancer incidence rate

According to cancer statistics, lung and bronchus cancer cause the most cancer deaths and have the second-largest incidence across cancer types in the USA (Siegel et al., 2019). In this study, cancer incidence data employed age group and sex classifications from SEER (NCI-DCCPS-SRP, 2022). Lung and bronchus cancer incidences between 2013 and 2017 of five-year age groups, spanning from 30 to 84 years old were used in this study. This age range was carefully selected to ensure a comprehensive analysis of lung cancer incidence across adulthood, capturing variations in risk that may emerge as individuals age. Further details on age classification can be found in the SEER*Stat documentation (NCI-DCCPS-SRP, 2022). These age groups were used in this study. Figure 7 shows cancer incidence rate for the 662 counties.

Regression models

Regression analysis was used to study the impact of various factors on health outcomes. Poisson regression, which is a count-based regression method, was utilized in this study.

The Poisson regression model used for the analysis is represented by the equation

$$\log\left(\lambda_i\right) = \alpha + \beta_1 \times e_1 + \beta_2 \times e_2 + \cdots + \beta_{\mathrm{ag1}} \times e_{\mathrm{ag1}} + \beta_{\mathrm{ag2}} \times e_{\mathrm{ag2}} + \cdots + \log(\mathrm{Pop}_i)$$

where λ_i is the expected count of the outcome variables for the ith observation, α is the intercept term, β_n denotes the coefficient for the nth predictor variables e_n is the values of the nth predictor variables and the log (Pop$_i$) is the offset term representing natural logarithm of the population for ith observation. For the age group variable, dummy variables were employed. $\beta_{\mathrm{ag}n}$ is the coefficient corresponding to the nth age group, and $e_{\mathrm{ag}n}$ is its associated variable. This dummy variable takes a value of 1 if the observation belongs to the nth age group, and 0 otherwise.

An RF approach was also employed by using rfPoisson from the R package fpechon/rfCountData (Liaw & Wiener, 2002; Pechon, 2019). The RF algorithm, which synthesizes the results from several simple trees of sequential specified questions or criteria to regress the data, can reduce the risk of overfitting on Poisson data (Pechon, 2019). The ML results were compared to Poisson regression through iterative fivefold cross-validation to evaluate the regression models. Comparisons were made with Mean Absolute Percentage Error (MAPE) (Hamner et al., 2018) and Root Mean Square Error (RMSE). In the first 5 times of fivefold cross-validation iteration, the RMSEs were computed. This was followed by a distinct 5 times of fivefold cross-validation process, in which the MAPEs and RMSEs were determined. Additionally, the variable importance measures (VIM, feature importance), which showed the importance of each factor that contributed to the regression results, was derived by using RF. VIMs were calculated using the '%IncLossFunction' metric from the random forest model, which measures the percentage increase in the model's loss function when the values of that feature are randomly permuted, indicating the significance of that feature in the model's predictive performance. Incidence rate ratios (IRRs) are reported only from the Poisson regression to increase the interpretability of the results.

MAPE

The regression results are evaluated using MAPE. This metric is scale-independent, which makes it possible to compare models across different datasets. A smaller MAPE indicates a better fitting model, where a value closer to 0 is preferred. However, it is sensitive to extreme values and positive errors. Additionally, if the actual value is close to 0, there is also a possibility that the error might be exaggerated even if the absolute error has a small value.

$$\mathrm{MAPE} = \mathrm{mean}\left(\left|\frac{g(x_t) - y_t}{y_t}\right|\right) * 100$$

where g is the regression model, and y_t is the target variable (de Myttenaere et al., 2016).

RMSE

RMSE was also used for evaluating the regression results. This metric is not a scale-independent, but one of the popular statistical metrics to be used to measure the magnitude of error between predicted and observed values.

$$\mathrm{RMSE} = \sqrt{\mathrm{mean}\left(\left(g(x_t) - y_t\right)^2\right)}$$

where g is the regression model, and y_t is the target variable.

MAPE and RMSE were calculated from a different validation process.

Table 2 Mean absolute percentage errors (MAPEs), root mean square errors (RMSEs), and their standard deviation from the test set with Poisson regression and Poisson random forest

ML or statistical models	Male		Female	
	MAPE (SD)	RMSE (SD)	MAPE (SD)	RMSE (SD)
Poisson regression	6.29 (2.67)	12.70 (3.94)	7.13 (4.04)	12.77 (3.38)
Poisson random forest	1.22 (0.0373)	8.01 (2.54)	1.16 (0.0391)	8.15 (2.52)

Table 3 Mean absolute percentage errors (MAPEs), root mean square errors (RMSEs), and their standard deviation of each data set with Poisson regression

Data	Male		Female	
	MAPE (SD)	RMSE (SD)	MAPE (SD)	RMSE (SD)
RadNet+radon	8.50 (2.67)	12.22 (3.94)	7.41 (3.83)	12.52 (3.86)
RadNet+radon+smoking+PM$_{2.5}$	7.76 (4.92)	12.75 (4.03)	20.13 (65.28)	12.66 (3.77)
RadNet+radon+smoking+PM$_{2.5}$+others	8.79 (10.10)	12.61 (4.25)	5.95 (1.92)	12.64 (3.57)
All (Full model)	6.29 (2.67)	12.70 (3.94)	7.13 (4.04)	12.77 (3.38)

Table 4 Mean absolute percentage errors (MAPEs), Root mean square errors (RMSEs), and their standard deviation of each data set with Random Forest

Data	Male		Female	
	MAPE (SD)	RMSE (SD)	MAPE (SD)	RMSE (SD)
RadNet+radon	1.42 (0.0603)	7.71 (2.36)	1.27 (0.0400)	7.84 (2.60)
RadNet+radon+smoking+$PM_{2.5}$	1.47 (0.0508)	9.21 (3.22)	1.51 (0.0323)	9.09 (3.21)
RadNet+radon+smoking+$PM_{2.5}$+others	1.36 (0.0390)	8.61 (3.24)	1.24 (0.0305)	8.57 (2.88)
All (Full model)	1.22 (0.0373)	8.01 (2.54)	1.16 (0.0391)	8.15 (2.52)

Table 5 Variable importance measures (VIM) of variables from male dataset with Random Forest

Variables	VIM
	Male
Age	8.10
Smoking	0.124
Median family income	$6.07*10^{-2}$
High school education	$4.86*10^{-3}$
Benzene	$4.84*10^{-3}$
Formaldehyde	$3.59*10^{-3}$
Unemployed	$7.52*10^{-4}$
RadNet	$5.26*10^{-4}$
Radon	$-3.05*10^{-3}$
$PM_{2.5}$	$-4.91*10^{-3}$
Urban	$-1.71*10^{-2}$
Ozone	$-2.23*10^{-2}$

Table 6 Variable importance measures (VIM) of variables from female dataset with Random Forest

Variables	VIM
	Female
Age	6.82
Formaldehyde	$2.29*10^{-2}$
Smoking	$1.61*10^{-2}$
Radon	$9.63*10^{-3}$
High school education	$7.11*10^{-3}$
Median family income	$4.50*10^{-3}$
RadNet	$4.08*10^{-3}$
Unemployed	$1.15*10^{-3}$
Benzene	$-1.91*10^{-3}$
$PM_{2.5}$	$-8.92*10^{-3}$
Ozone	$-2.89*10^{-2}$
Urban	-0.158

Results

By developing a dataset of radiation, environmental, and sociodemographic variables that span the period of 2013–2017 (Table 1), Poisson regression and Poisson RF models were employed to model the relationship between the cancer-related factors and the lung/bronchus cancer incidence.

MAPE showed statistically significant differences when *T*-test was done between Poisson regression and Poisson RF. As the number of samples for each case is 25, degree of freedom is 48. For both males ($t(48) = 12.86$, $p < 0.01$) and females ($t(48) = 6.40$, $p < 0.01$). RMSE also showed significant differences for both males ($t(48) = 8.85$, $p < 0.01$) and females ($t(48) = 6.57$, $p < 0.01$) (Table 2).

Tables 3 and 4 summarize the regression results of various datasets through Poisson RF and Poisson regression. Smoking, radiation exposure, and $PM_{2.5}$, which are thought to be related to radon exposure (Matthaios et al., 2021; Trassierra et al., 2016), and sociodemographic and behavioral factors were combined in various models. The analysis of the relationship between variables and model accuracy revealed an interesting trend in the error from the Poisson RF, as shown in Table 4. The VIM was acquired by averaging the model weights across folds with the entire dataset by using the default function of the fpechon/rfCountData package (Liaw & Wiener, 2002; Pechon, 2019). Table 5 and 6 show the VIMs of the variables analyzed with full model Poisson random forest regression including all variables, including socioeconomic variables, in the model.

Table 7 summarizes the IRRs analyzed with full model Poisson regression. The increased unit of the IRR is proportional to the range of each variable to make a more intuitive comparison. In both

Table 7 Incidence rate ratios (IRRs) and 95% confidence intervals of each factor of interest with poisson regression

	IRR		Unit of measurement
	Male (95% CI)	Female (95% CI)	
Smoking	1.47 (1.45,1.48)	1.45 (1.44, 1.47)	Per 5% increase in the population
Radon	0.99 (0.98, 0.99)	0.99 (0.98, 0.99)	Per 3 pCi/L increase
RadNet	0.97 (0.97, 0.98)	0.98 (0.98, 0.99)	Per 500 cpm increase
$PM_{2.5}$	1.02 (1.01, 1.03)	0.99 (0.98, 1.00)	Per 3 μg/m^3 increase
Formaldehyde	1.02 (1.02, 1.02)	0.98 (0.98, 0.99)	Per 0.3 μg/m^3 increase
Benzene	1.01 (1.00, 1.02)	1.02 (1.01, 1.02)	Per 0.3 μg/m^3 increase

Table 8 Incidence rate ratios (IRRs) and 95% confidence intervals of $PM_{2.5}$ and smoking by radon zone

	Radon zone 1		Radon zone 2		Radon zone 3	
	Male (95% CI)	Female (95% CI)	Male (95% CI)	Female (95% CI)	Male (95% CI)	Female (95% CI)
IRR of $PM_{2.5}$	1.17 (1.13, 1.20)	1.13 (1.10, 1.17)	1.04 (1.03, 1.06)	1.03 (1.01, 1.05)	0.92 (0.89, 0.94)	0.94 (0.91, 0.96)
IRR of smoking	1.36 (1.31, 1.40)	1.45 (1.41, 1.50)	1.52 (1.49, 1.54)	1.55 (1.52, 1.58)	1.50 (1.47, 1.53)	1.35 (1.32, 1.37)

cases, smoking had the greatest effect on lung cancer incidence rates. In the case of indoor radon, the association was negative. [Male: 0.99 (0.98, 0.99), Female: 0.99 (0.98, 0.99)]. Also, Background gamma count (RadNet) [Male: 0.97 (0.97, 0.98), Female: 0.98 (0.98, 0.99)] and three-year average $PM_{2.5}$ for female [0.99 (0.98, 1.00) *P*-value: 0.09] showed negative associations at higher concentrations, which somewhat contradicts results from previous studies (Ghazipura et al., 2019; Raaschou-Nielsen et al., 2013; Turner et al., 2011a, 2011b).

To understand the differences broken down by EPA Radon Zone, separate regression models were run for each zone using the full model Poisson regression (Table 8). In the case of Radon Zone 1, an area with high radon concentration, the effect of $PM_{2.5}$ exposure was the greatest. Conversely, in the case of Radon Zone 3, which is an area with a low radon concentration, higher rates of $PM_{2.5}$ were associated with lower incidence rates. The effect of smoking was consistent across all radon zones.

Discussion

The effects of environmental exposure on health outcomes are complex. In this study, the results (Table 8) suggest that the assocation between $PM_{2.5}$ may vary with levels of indoor radon exposure. Despite potential synergistic effects of exposure, many radiation epidemiological studies include a limited number of environmental exposure measures (Haylock et al., 2018; Richardson et al., 2015; Stanley et al., 2019; Tomasek, 2013). Belloni et al. (2020) have noted that few studies (Klebe et al., 2019; Leuraud et al., 2011) have attempted to address multifactorial exposures from environmental stressors. In the study of radiation-related disease, estimating the risk associated with radiation-related lung cancer has been a focal point in resolving the dose-risk response relationship (United Nations Scientific Committee on the Effects of Atomic Radiation [UNSCEAR], 2018). Furthermore, due to the high baseline cancer risk compared to the risk increased from low-dose radiation exposure, the population size required for detecting low-dose radiation risk with statistical significance exponentially increases as the target dose decreases (Ozasa, 2016; Ozasa et al., 2019; UNSCEAR, 2008; Valentin, 2006). To address some of the challenges, studies that use a wider range of data, such as the Million Person Study (Boice et al., 2022), are being conducted (Calabrese, 2015; Ricci & Tharmalingam, 2019; Tubiana et al., 2009; Valentin, 2008; Weber & Zanzonico, 2017). The utilization of population-level exposure variables and health outcomes data adopted in this study can serve as a valuable resource for future research. Population-level data offers an advantage in the adoption of multiple variables and the analysis of diverse health outcomes. Furthermore, ML techniques are particularly well suited to model the complex relationships that exist between environmental exposure and health outcomes. By leveraging ML, it is possible to capture the complex interplay between environmental exposures and health, thereby offering a promising avenue for future research in this field.

The results suggest that $PM_{2.5}$ should be included in future analysis of radon-induced lung cancer incidence, as there may be an interaction with radon exposure. The observed patterns, where changes in radon concentration result in significant differences ($p < 0.001$ for all cases) in the effects of $PM_{2.5}$, corroborate findings from other research that explores the combined impacts of $PM_{2.5}$ and radon exposure (Dlugosz-Lisiecka, 2016). $PM_{2.5}$ or other particulate matter could be one of the possible transport mechanisms that allow radon gas to permeate lung tissue. This is further supported by two experimental studies that assess the speciation of $PM_{2.5}$ particles in the presence of radon progeny. The first study shows that the alpha activity of $PM_{2.5}$ tends to increase as the concentration of radon increases (Matthaios et al., 2021). The second study shows that in a radon chamber, the presence of particulate matter will increase the attached fraction of radon progeny, thereby implying that the radiation exposure from particulate matter will increase (Trassierra et al., 2016). $PM_{2.5}$ and radon seem to have synergistic effects and are thought to affect various health outcomes, including incidences of lung cancer. Given the possible synergistic effect between $PM_{2.5}$ and radon, future epidemiological studies should investigate this further.

This study harnessed ML to consider the nonlinear effects of radon exposure within the context of other environmental factors. The results of decreased errors from ML models show that ML is effective at analyzing complex relationships in environmental exposure studies and should be considered in future studies that investigate the relationship between radon exposure and cancer outcomes. One limitation of current ML is the lack of variety in ML algorithm packages that can be applied to count data. However, it is believed that these problems will naturally be resolved as ML develops and becomes more widely used in regression analysis.

Large-scale data can be challenging when conducting analysis attributable to individual characteristics, for example they are limited in their ability to reflect the interaction of environmental and genomic factors, which is important in the exposome approach (Zhang et al., 2021). Furthermore, individual history of exposure information which is similarly essential to exposome analysis is difficult to reflect in the analysis (Zhang et al., 2021). Thus, population-level studies of incidence rates, such as this one, are susceptible to the ecological fallacy. This limits the ability to establish causal relationships between variables and health outcomes. Despite these limitations, population-level studies can still provide valuable reference points for guiding individual-level studies.

The World Health Organization (2009) reported that radon is the second major contributor to lung cancer incidence. Also, a study by Turner et al. (2011a, 2011b), which analyzed county-level radon concentrations and residents' lung cancer risk similar to this study, showed a positive association between residential radon and lung cancer risk. However, our results showed that there was negative association between radon and lung cancer incidence rates [IRR of male: 0.99 (0.98, 0.99), IRR of female: 0.99 (0.98, 0.99)]. There are several reasons our findings may differ from occupational cohort studies that show there is a strong association in occupational studies where individuals are exposed to high levels radon (Kreuzer et al., 2015; Leuraud et al., 2011; Richardson et al., 2021, 2022). First, as mentioned above, this study may suffer from ecological fallacy. Second, indoor radon exposure risk is measured at the county level and radon exposure varies widely across counties (Li et al., 2021). Third, the effect sizes at low levels of exposure are likely small—making the signal difficult to detect in an ecological analysis. Our results of study which investigated the association between residential radon exposure to lung cancer is difficult distinguished are more aligned with results from recently published residential exposure and lung cancer-based study (Li et al., 2020). The study on residential radon exposure and lung cancer risk in Connecticut and Utah (Sandler et al., 2006) could not provide evidence of an increased risk of lung cancer at the exposure levels observed. Unlike minor studies, the residential radon exposure is so low that statistically significant results are difficult to obtain.

Furthermore, the difference in findings across studies may arise from discrepancies between individual-level and population-level approaches in their methodologies and analysis. Also, regarding the interaction between smoking and radon, the results were different from the previous studies. According to BEIR VI, a comprehensive analysis of the relationship between smoking habits, radon exposure, and lung cancer risk of uranium miners from several studies showed a submultiplicative effect, which means that the risk in the population exposed to both smoking and radon is greater than the sum of the individual risks expected from either smoking or radon exposure and less than the product (NRC, 1999). The results of a case-control study in Spain after BEIR VI indicated that there is a strong synergistic effect between smoking and radon exposure, and the case-control miner study showed evidence of submultiplicative interaction between radon and smoking (Barros-Dios et al., 2012; Leuraud et al., 2011). However, the association between smoking and radon concentration did not appear to be significant in the results presented herein. These inconsistent results again may be attributed to certain limitations in this study, including terse measurement of radon concentrations. Using the median data could prevent the effects of outliers, but it will have errors from the insufficient number of tests. This problem could skew the results toward non-significant associations or even contradict established knowledge.

Possible confounding factors that were not properly reflected are that the level of stress that people experience, and the quality of medical care will vary considerably by county or state despite some socioeconomic factors being included. This may also explain the opposite trend in this analysis vs. the previously known results of $PM_{2.5}$ and lung cancer incidence. These problems could be mitigated if the research is conducted on specific regions with very high-resolution data, or by improving our measures of radon concentrations. Another limitation of this study is the lack of residential history data, which made it impossible to create a model that adequately considers different exposures across a life span and the associated latency periods. Other lung cancer models have considered the incubation period of 5 years (National Research Council [NRC], 2006; UNSCEAR, 2008; Valentin, 2008). Future studies should use residential history to assess the effects of indoor radon exposure across a life span.

If future studies address these limitations, then the combination of highly accurate ML techniques and the advantages and applicability in radiation epidemiology of population-level data could be harnessed for more diverse health outcome analysis. This may also provide valuable insights into the interplay between variables.

Conclusion

Traditional statistical methods and ML models can be used in parallel to fully understand the complex relationship between environmental exposures and health. To investigate the applicability of multivariable and ML methods in environmental exposure studies, county-level lung/bronchus cancer risk was assessed with various exposures (airborne gamma counts, radon concentration, air quality), lifestyle (smoking), and socioeconomic factors through Poisson regression and Poisson RF regression. The study found that the risk of lung cancer from $PM_{2.5}$ varied by radon concentration with larger effect sizes in areas with high indoor radon exposure. In summary, the results of this study demonstrate how (1) including multiple environmental exposures has advantages over single exposure studies when the relationship between the environment and lung cancer risk is considered, thereby making an exposomics framework an important consideration, and (2) employing ML models enhances the utility of analysis in identifying complex relationships, as in the case of environmental radiation exposure and lung cancer incidence. Consequently, this study proposes a new paradigm for studying environmental radiation combined with other environmental exposures.

Acknowledgements This work was supported by the Office of Biological and Environmental Research's Biological Systems Science Division. This manuscript has been authored by UT-Battelle LLC under Contract No. DE-AC05-00OR22725 with the US Department of Energy and Award AWD-002827 between UT-Battelle and the Georgia Tech Research Corporation.

Author contributions HL wrote the main manuscript text, collected, processed, and analyzed the data, and contributed to the interpretation of results. HH, GA, and SD assisted in the design and implementation of the study and interpretation of the results. JL contributed the most to data collection, processing, and visualization. DM contributed to the verification of code compliance. SD, HH, GA, and AK supervised the overall project. All authors reviewed the manuscript for intellectual content and provided critical feedback.

Funding Funding was provided by The Office of Biological and Environmental Research's Biological Systems Science Division.

Declarations

Conflict of interest The authors declare no potential conflicts of interest**.**

References

Abergel, R., Aris, J., Bolch, W. E., Dewji, S. A., Golden, A., Hooper, D. A., Margot, D., Menker, C. G., Paunesku, T., Schaue, D., & Woloschak, G. E. (2022). The enduring legacy of Marie Curie: Impacts of radium in 21st century radiological and medical sciences. *International Journal of Radiation Biology, 98*(3), 267–275. https://doi.org/10.1080/09553002.2022.2027542

Barros-Dios, J. M., Ruano-Ravina, A., Perez-Rios, M., Castro-Bernardez, M., Abal-Arca, J., & Tojo-Castro, M. (2012). Residential radon exposure, histologic types, and lung cancer risk. A case-control study in Galicia Spain. *Cancer Epidemiol Biomarkers & Prevention, 21*(6), 951–958. https://doi.org/10.1158/1055-9965.EPI-12-0146-T

Belloni, M., Laurent, O., Guihenneuc, C., & Ancelet, S. (2020). Bayesian profile regression to deal with multiple highly correlated exposures and a censored survival outcome. First application in ionizing radiation epidemiology. *Frontiers in Public Health, 8*, 557006. https://doi.org/10.3389/fpubh.2020.557006

Blomberg, A. J., Coull, B. A., Jhun, I., Vieira, C. L. Z., Zanobetti, A., Garshick, E., Schwartz, J., & Koutrakis, P. (2019). Effect modification of ambient particle mortality by radon: a time series analysis in 108 US cities. *Journal of the Air & Waste Management Association, 69*(3), 266–276. https://doi.org/10.1080/10962247.2018.1523071

Boice, J. D., Jr., Cohen, S. S., Mumma, M. T., & Ellis, E. D. (2022). the million person study, whence it came and why. *International Journal of Radiation Biology, 98*(4), 537–550. https://doi.org/10.1080/09553002.2019.1589015

Calabrese, E. J. (2015). Model uncertainty via the integration of hormesis and LNT as the default in cancer risk assessment. *Dose Response, 13*(4), 1559325815621764. https://doi.org/10.1177/1559325815621764

Centers for Disease Control and Prevention. (n.d.). *National Environmental Public Health Tracking Network*. Retrieved 8 Aug 2022, from https://ephtracking.cdc.gov/DataExplorer/.

Cohen, B. L. (1995). Test of the linear-no threshold theory of radiation carcinogenesis for inhaled radon decay products. *Health Physics, 68*(2), 157–174.

Cohen, B. L., & Colditz, G. A. (1994). Tests of the linear-no threshold theory for lung cancer induced by exposure to radon. *Environmental Research, 64*(1), 65–89.

Couraud, S., Zalcman, G., Milleron, B., Morin, F., & Souquet, P. J. (2012). Lung cancer in never smokers-A review. *European Journal of Cancer, 48*(9), 1299–1311. https://doi.org/10.1016/j.ejca.2012.03.007

de Groot, P. M., Wu, C. C., Carter, B. W., & Munden, R. F. (2018). The epidemiology of lung cancer. *Translational Lung Cancer Research, 7*(3), 220–233. https://doi.org/10.21037/tlcr.2018.05.06

de Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing, 192*, 38–48. https://doi.org/10.1016/j.neucom.2015.12.114

Dela Cruz, C. S., Tanoue, L. T., & Matthay, R. A. (2011). Lung cancer: Epidemiology, etiology, and prevention. *Clinics in Chest Medicine, 32*(4), 605–644. https://doi.org/10.1016/j.ccm.2011.09.001

Dlugosz-Lisiecka, M. (2016). The sources and fate of (210)Po in the urban air: A review. *Environment International, 94*, 325–330. https://doi.org/10.1016/j.envint.2016.06.002

Dong, S., Koutrakis, P., Li, L., Coull, B. A., Schwartz, J., Kosheleva, A., & Zanobetti, A. (2022). Synergistic effects of particle radioactivity (gross β activity) and particulate matter ≤ 2.5 mum aerodynamic diameter on cardiovascular disease mortality. *Journal of the American Heart Association, 11*(20), e025470. https://doi.org/10.1161/JAHA.121.025470

Fraass, R. (2015). RadNet national air monitoring program. In S. Apikyan & D. Diamond (Eds.), *Nuclear terrorism and national preparedness* (pp. 117–123). Springer.

Ghazipura, M., Garshick, E., & Cromar, K. (2019). Ambient $PM_{2.5}$ exposure and risk of lung cancer incidence in North America and Europe. *Environmental Research Communications, 1*(1), 015004. https://doi.org/10.1088/2515-7620/ab06e9

Hamner, B., Frasco, M., & LeDell, E. (2018). Metrics: Evaluation metrics for machine learning. *R package version 0.1, 4*.

Haylock, R. G. E., Gillies, M., Hunter, N., Zhang, W., & Phillipson, M. (2018). Cancer mortality and incidence following external occupational radiation exposure: An update of the 3rd analysis of the UK national registry for radiation workers. *British Journal of Cancer, 119*(5), 631–637. https://doi.org/10.1038/s41416-018-0184-9

Klebe, S., Leigh, J., Henderson, D. W., & Nurminen, M. (2019). Asbestos, smoking and lung cancer: An update. *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph17010258

Kreuzer, M., Fenske, N., Schnelzer, M., & Walsh, L. (2015). Lung cancer risk at low radon exposure rates in German uranium miners. *British Journal of Cancer, 113*(9), 1367–1369. https://doi.org/10.1038/bjc.2015.324

Kreuzer, M., Grosche, B., Schnelzer, M., Tschense, A., Dufey, F., & Walsh, L. (2010). Radon and risk of death from cancer and cardiovascular diseases in the German uranium miners cohort study: Follow-up 1946–2003. *Radiation and Environmental Biophysics, 49*(2), 177–185. https://doi.org/10.1007/s00411-009-0249-5

Leuraud, K., Schnelzer, M., Tomasek, L., Hunter, N., Timarche, M., Grosche, B., Kreuzer, M., & Laurier, D. (2011). Radon, smoking and lung cancer risk: Results of a joint analysis of three European case-control studies among uranium miners. *Radiation Research, 176*(3), 375–387. https://doi.org/10.1667/rr2377.1

Li, C., Wang, C., Yu, J., Fan, Y., Liu, D., Zhou, W., & Shi, T. (2020). Residential radon and histological types of lung cancer: A meta-analysis of case-control studies. *International Journal of Environmental Research and Public Health, 17*(4), 1457. https://doi.org/10.3390/ijerph17041457

Li, L., Blomberg, A. J., Stern, R. A., Kang, C. M., Papatheodorou, S., Wei, Y., & Koutrakis, P. (2021). Predicting monthly community-level domestic radon concentrations in the greater Boston area with an ensemble learning model. *Environmental Science & Technology, 55*(10), 7157–7166. https://doi.org/10.1021/acs.est.0c08792

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22.

Lyon, J. L., Alder, S. C., Stone, M. B., Scholl, A., Reading, J. C., Holubkov, R., Sheng, X., White, G. L., Jr., Hegmann, K. T., Anspaugh, L., Hoffman, F. O., Simon, S. L., Thomas, B., Carroll, R., & Meikle, A. W. (2006). Thyroid disease associated with exposure to the nevada nuclear weapons test site radiation: A reevaluation based on corrected dosimetry and examination data. *Epidemiology, 17*(6), 604–614. https://doi.org/10.1097/01.ede.0000240540.79983.7f

Matthaios, V. N., Liu, M., Li, L., Kang, C.-M., Vieira, C. L., Gold, D. R., & Koutrakis, P. (2021). Sources of indoor $PM_{2.5}$ gross α and β activities measured in 340 homes. *Environmental Research, 197*, 111114.

McDonald, J. W., Taylor, J. A., Watson, M. A., Saccomanno, G., & Devereux, T. R. (1995). p53 and K-ras in radon-associated lung adenocarcinoma. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology, 4*(7), 791–793.

Mifune, M., Sobue, T., Arimoto, H., Komoto, Y., Kondo, S., & Tanooka, H. (1992). Cancer mortality survey in a spa area (Misasa, Japan) with a high radon background. *Japanese Journal of Cancer Research, 83*(1), 1–5.

Moritz, S., & Bartz-Beielstein, T. (2017). imputets: Time series missing value imputation in R. *R J, 9*(1), 207.

Morton, L. M., Karyadi, D. M., Stewart, C., Bogdanova, T. I., Dawson, E. T., Steinberg, M. K., Dai, J., Hartley, S. W., Schonfeld, S. J., Sampson, J. N., Maruvka, Y. E., Kapoor, V., Ramsden, D. A., Carvajal-Garcia, J., Perou, C. M., Parker, J. S., Krznaric, M., Yeager, M., Boland, J. F., & Chanock, S. J. (2021). Radiation-related genomic profile of papillary thyroid carcinoma after the Chernobyl accident. *Science*. https://doi.org/10.1126/science.abg2538

Mose, D. G., & Mushrush, G. W. (1999). Prediction of indoor radon based on soil radon and soil permeability. *Journal of Environmental Science & Health Part A, 34*(6), 1253–1266.

National Research Council. (1999). *Health effects of exposure to radon: BEIR VI*.

National Research Council. (2006). *Health risks from exposure to low levels of ionizing radiation: BEIR VII phase 2*.

National Cancer Institute, DCCPS, Surveillance Research Program. (2022). SEER*Stat database: Incidence-SEER research data, 8 registries, Nov 2021 Sub (1975–2019)-linked to county attributes-time dependent (1990–2019) income/rurality, 1969–2020 counties. Released April 2022, based on the November 2021 submission. https://www.seer.cancer.gov.

Ozasa, K. (2016). Epidemiological research on radiation-induced cancer in atomic bomb survivors. *Journal of Radiation Research, 57*(S1), i112–i117. https://doi.org/10.1093/jrr/rrw005

Ozasa, K., Cullings, H. M., Ohishi, W., Hida, A., & Grant, E. J. (2019). Epidemiological studies of atomic bomb radiation at the radiation effects research foundation. *International Journal of Radiation Biology, 95*(7), 879–891. https://doi.org/10.1080/09553002.2019.1569778

Pechon, F. (2019). rfCountData [R package]. GitHub. Documentation built on August 12, 2019, 11:16 a.m. Retrieved 8 Aug 2022, from https://rdrr.io/github/fpechon/rfCountData.

Ponciano-Rodríguez, G., Gaso, M., Armienta, M., Trueta, C., Morales, I., Alfaro, R., & Segovia, N. (2021). Indoor radon exposure and excess of lung cancer mortality: The case of Mexico—An ecological study. *Environmental Geochemistry and Health, 43*, 221–234.

Przylibski, T. A., Staśko, S., & Domin, E. (2022). Radon groundwater in a radon-prone area: Possible uses and problems: An example from SW part of Kłodzko Valley, Sudetes SW Poland. *Environmental Geochemistry and Health, 44*(12), 4539–4555.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raaschou-Nielsen, O., Andersen, Z. J., Beelen, R., Samoli, E., Stafoggia, M., Weinmayr, G., Hoffmann, B., Fischer, P., Nieuwenhuijsen, M. J., Brunekreef, B., Xun, W. W., Katsouyanni, K., Dimakopoulou, K., Sommar, J., Forsberg, B., Modig, L., Oudin, A., Oftedal, B., Schwarze, P. E., & Hoek, G. (2013). Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). *The Lancet Oncology, 14*(9), 813–822. https://doi.org/10.1016/S1470-2045(13)70279-1

Remington, P. L., Catlin, B. B., & Gennuso, K. P. (2015). The County Health Rankings: Rationale and methods. *Population Health Metrics, 13*, 11. https://doi.org/10.1186/s12963-015-0044-2

Ricci, P. F., & Tharmalingam, S. (2019). Ionizing radiations epidemiology does not support the LNT model.

Chemico-Biological Interactions, 301, 128–140. https://doi.org/10.1016/j.cbi.2018.11.014

Richardson, D. B., Cardis, E., Daniels, R. D., Gillies, M., O'Hagan, J. A., Hamra, G. B., Haylock, R., Laurier, D., Leuraud, K., Moissonnier, M., Schubauer-Berigan, M. K., Thierry-Chef, I., & Kesminiene, A. (2015). Risk of cancer from occupational exposure to ionising radiation: Retrospective cohort study of workers in France, the United Kingdom, and the United States (INWORKS). *BMJ, 351*, h5359. https://doi.org/10.1136/bmj.h5359

Richardson, D. B., Rage, E., Demers, P. A., Do, M. T., DeBono, N., Fenske, N., Deffner, V., Kreuzer, M., Samet, J., Wiggins, C., Schubauer-Berigan, M. K., Kelly-Reif, K., Tomasek, L., Zablotska, L. B., & Laurier, D. (2021). Mortality among uranium miners in North America and Europe: The Pooled Uranium Miners Analysis (PUMA). *International Journal of Epidemiology, 50*(2), 633–643. https://doi.org/10.1093/ije/dyaa195

Richardson, D. B., Rage, E., Demers, P. A., Do, M. T., Fenske, N., Deffner, V., Kreuzer, M., Samet, J., Bertke, S. J., Kelly-Reif, K., Schubauer-Berigan, M. K., Tomasek, L., Zablotska, L. B., Wiggins, C., & Laurier, D. (2022). Lung cancer and radon: Pooled analysis of uranium miners hired in 1960 or later. *Environmental Health Perspectives, 130*(5), 57010. https://doi.org/10.1289/EHP10669

Sandler, D. P., Weinberg, C. R., Shore, D. L., Archer, V. E., Stone, M. B., Lyon, J. L., Rothney-Kozlak, L., Shepherd, M., & Stolwijk, J. A. (2006). Indoor radon and lung cancer risk in connecticut and utah. *Journal of Toxicology and Environmental Health. Part A, 69*(7), 633–654. https://doi.org/10.1080/15287390500261117

Seong, K. M., Seo, S., Lee, D., Kim, M. J., Lee, S. S., Park, S., & Jin, Y. W. (2016). Is the linear no-threshold dose-response paradigm still necessary for the assessment of health effects of low dose radiation? *Journal of Korean Medical Science, 31*(Suppl 1), S10-23. https://doi.org/10.3346/jkms.2016.31.S1.S10

Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians, 69*(1), 7–34. https://doi.org/10.3322/caac.21551

Smith, B. J., & Field, R. W. (2007). Effect of housing factors and surficial uranium on the spatial prediction of residential radon in Iowa. *Environmetrics: The Official Journal of the International Environmetrics Society, 18*(5), 481–497.

Stanley, F. K., Irvine, J. L., Jacques, W. R., Salgia, S. R., Innes, D. G., Winquist, B. D., Torr, D., Brenner, D. R., & Goodarzi, A. A. (2019). Radon exposure is rising steadily within the modern North American residential environment, and is increasingly uniform across seasons. *Scientific Reports, 9*(1), 1–17.

Takahashi, T., Schoemaker, M., Trott, K., Simon, S., Fujimori, K., Nakashima, N., Fukao, A., & Saito, H. (2003). The relationship of thyroid cancer with radiation exposure from nuclear weapon testing in the Marshall Islands. *Journal of Epidemiology, 13*(2), 99–107.

Tirmarche, M., Harrison, J., Laurier, D., Paquet, F., Blanchardon, E., & Marsh, J. (2010). ICRP Publication 115. Lung cancer risk from radon and progeny and statement on radon. *Annals of the ICRP, 40*(1), 1–64.

Tomasek, L. (2013). Lung cancer risk from occupational and environmental radon and role of smoking in two Czech nested case-control studies. *International Journal of Environmental Research and Public Health, 10*(3), 963–979. https://doi.org/10.3390/ijerph10030963

Trassierra, C. V., Stabile, L., Cardellini, F., Morawska, L., & Buonanno, G. (2016). Effect of indoor-generated airborne particles on radon progeny dynamics. *Journal of Hazardous Materials, 314*, 155–163. https://doi.org/10.1016/j.jhazmat.2016.04.051

Tubiana, M., Feinendegen, L. E., Yang, C., & Kaminski, J. M. (2009). The linear no-threshold relationship is inconsistent with radiation biologic and experimental data. *Radiology, 251*(1), 13.

Turner, M. C., Krewski, D., Chen, Y., Pope, C. A., 3rd., Gapstur, S., & Thun, M. J. (2011a). Radon and lung cancer in the American cancer society cohort. *Cancer Epidemiology, Biomarkers & Prevention, 20*(3), 438–448. https://doi.org/10.1158/1055-9965.EPI-10-1153

Turner, M. C., Krewski, D., Pope, C. A., 3rd., Chen, Y., Gapstur, S. M., & Thun, M. J. (2011b). Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers. *American Journal of Respiratory and Critical Care Medicine, 184*(12), 1374–1381. https://doi.org/10.1164/rccm.201106-1011OC

United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR). (2008). *Effects of Ionizing Radiation: 2006 Report, Volume I: Report to the General Assembly, Scientific Annexes A and B*. United Nations.

United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR). (2018). *Sources, Effects and Risks of Ionizing Radiation: 2017 Report*. United Nations. https://doi.org/10.18356/7e4f1c5a-en

University of Wisconsin Population Health Institute. (2022). *County Health Rankings & Roadmaps*. Retrieved 8 Aug 2022, from https://www.countyhealthrankings.org.

US Environmental Protection Agency. (1993). *EPA's map of radon zones national summary*. US Environmental Protection Agency.

US Environmental Protection Agency. (n.d.). RadNet. Retrieved 5 May 2022, from https://www.epa.gov/radnet/.

Valentin, J. (2006). *Low-dose extrapolation of radiation-related cancer risk*. Elsevier International Commission on Radiological Protection.

Valentin, J. (2008). *The 2007 recommendations of the international commission on radiological protection*. Elsevier International Commission on Radiological Protection.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., & SciPy, C. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Natural Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Weber, W., & Zanzonico, P. (2017). The Controversial Linear No-Threshold Model. *Journal of Nuclear Medicine, 58*(1), 7–8. https://doi.org/10.2967/jnumed.116.182667

Wiemken, T. L., & Kelley, R. R. (2020). Machine learning in epidemiology and health outcomes research. *Annual*

Review of Public Health, 41, 21–36. https://doi.org/10.1146/annurev-publhealth-040119-094437

Wild, C. P. (2005). Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology, Biomarkers & Prevention, 14*(8), 1847–1850. https://doi.org/10.1158/1055-9965.EPI-05-0456

World Health Organization. (2009). *WHO handbook on indoor radon: A public health perspective*.

Yoon, J. Y., Lee, J. D., Joo, S. W., & Kang, D. R. (2016). Indoor radon exposure and lung cancer: A review of ecological studies. *Annals of Occupational and Environmental Medicine, 28*, 15. https://doi.org/10.1186/s40557-016-0098-z

Zhang, P., Carlsten, C., Chaleckis, R., Hanhineva, K., Huang, M., Isobe, T., Koistinen, V. M., Meister, I., Papazian, S., Sdougkou, K., Xie, H., Martin, J. W., Rappaport, S. M., Tsugawa, H., Walker, D. I., Woodruff, T. J., Wright, R. O., & Wheelock, C. E. (2021). Defining the scope of exposome studies and research needs from a multidisciplinary perspective. *Environmental Science & Technology Letters, 8*(10), 839–852. https://doi.org/10.1021/acs.estlett.1c00648

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.