Published in final edited form as:

Spat Spatiotemporal Epidemiol. 2020 November ; 35: 100364. doi:10.1016/j.sste.2020.100364.

Using residential histories in case-control analysis of lung cancer and mountaintop removal coal mining in Central Appalachia

WJ Christian, PhD, MPH^a, CJ Walker, BA^a, B Huang, DrPH, MS^b, JE Levy, MPH^c, E Durbin, DrPH, MS^b, S Arnold, MD^d

^aDept. of Epidemiology College of Public Health, University of Kentucky, Lexington, Kentucky, USA

^bKentucky Cancer Registry, Lexington, Kentucky, USA.

^cDept. of Geography, College of Arts & Sciences, University of Kentucky, Lexington, Kentucky, USA.

^dMarkey Cancer Center, University of Kentucky, Lexington, Kentucky, USA.

Abstract

Population-based ecological and cross-sectional studies have observed high risk for several cancers in areas of Central Appalachia where mountaintop removal coal mines operate. Case-control studies could provide stronger evidence of such relationships, but misclassification of exposure is likely when based on current residence, since individuals could have inhabited several residences with varying environmental exposures over many years. To address this, we used residential histories for individuals enrolled in a previous case-control study of lung cancer to assess residential proximity to mountaintop removal coal mining over a 30-year period, using both survey data and proprietary data from LexisNexis, Inc. Supplementing the survey data with LexisNexis data improved precision and completeness of geographic coordinates. Final logistic regression models revealed higher odds of high exposure among cases. These findings suggest that living in close proximity to mountaintop removal coal mining sites could increase risk for lung cancer, after adjusting for other relevant factors.

Keywords

Residential history; Lung cancer; LexisNexis; Environment; Appalachia; Exposure assessment; Cancer

Corresponding Author: W. Jay Christian, PhD, MPH, Assistant Professor, Dept. of Epidemiology, College of Public Health, University of Kentucky, 111 Washington Ave, #211B, Lexington, KY 40536, jay.christian@uky.edu, Tel: (859) 218-2016.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of interests

The authors have no financial disclosures to declare and no conflicts of interest to report.

1. Introduction

Multiple ecological and cross-sectional studies published over the past two decades have demonstrated that risk for multiple diseases—including many types of cancer, birth defects, respiratory disease, and kidney disease—are associated with residence near mountaintop removal (MTR) coal mining sites.^{1–3} Briefly, MTR coal mining is a type of surface mining where coal seams are exposed by removing vegetation, rock, and soil (overburden) from the top of a mountain.⁴ Previous research has shown that lung cancer rates are elevated among populations in such coal mining regions, even after adjusting for rates of cigarette smoking, the foremost cause of the disease.^{5–7} Few rigorous individual-level studies of lung cancer risk and MTR coal mining have been reported, however, despite the wealth of population-based evidence. Case-control study designs in particular have the potential to provide stronger evidence of a relationship, if one truly exists. One recent study, however, conducted in eastern Kentucky and published by Unrine and colleagues in 2019, failed to identify a significant association between lung cancer and trace elements exposure.⁵ In that study, cases' and controls' toenails had similar levels of most trace elements, and very few participants' toenails had high concentrations of any trace elements.

Latency, however, is a major challenge for case-control studies of environmental exposures and cancer. Due to the decades-long latency of lung cancer, estimating exposures occurring over decades is sometimes necessary to accurately calculate risk. It is possible, given the continuing decline of coal mining in eastern Kentucky,⁸ that current lung cancer rates are elevated due to environmental exposures that have substantially, recently, and rapidly abated. Residential histories with sufficient spatial and temporal detail can provide spatial data for long-term exposure assessment, but their quality is often hindered by participants' difficulty with recall of particular details, especially full street addresses including house or building number. Without such precise locational data, it is much more difficult to accurately assess potential exposures among research participants, since most geocoding algorithms will assign individuals to locations such as ZIP code or county centroids (when lacking a recognizable street address), and this could be miles from their actual residence, especially in rural regions.⁹ This is precisely the case with residential history data collected by the study described in Unrine et al. for all 150 cases and 370 controls who participated, as will be shown.⁵

Supplementing currently existing residential histories with address data from LexisNexis, Inc. (LN), a commercial provider of data services, has previously demonstrated promise for robust analysis of cancer risk in relation to residential history. Jacquez and colleagues conducted an analysis of bladder cancer in coordination with the Michigan Cancer Registry, and Hurley and colleagues more recently completed an analysis of breast cancer using data from the California Teachers Study (CTS).^{6,7} Both studies demonstrated that address data from LexisNexis provided information that further supplemented the residential history survey data already collected, but also presented some notable challenges, including duplicate records and invalid dates of residence. This source of addresses has inherent flaws, like surveys, but appears to contribute useful information for exposure assessment in environmental epidemiologic studies of cancer. Still, these innovative techniques have not yet been employed in eastern Kentucky or elsewhere in the Central Appalachian region,

which has both a legacy of resource extraction and very high rates of cancer and other diseases known to have environmental etiologies. Additionally, other researcher teams have developed algorithms to process address data from LN, including Westat Inc.,¹⁰ but a notable limitation is they do not allow for overlapping times of residence at multiple locations, nor residence at the same address during non-adjacent time periods. Wheeler and Wang, using a cohort of participants from the American Association of Retired Persons (AARP), found that LN data could be useful for reconstructing residential histories, although match rates declined in earlier years, and U.S. regions had varying match rates.¹¹

The goal of this study was to use address data gathered from surveys and LexisNexis to create residential histories for use in a case-control analysis of lung cancer and MTR mining in Central Appalachia. Using the same research participants included in the study by Unrine and colleagues, we examined the relationship between lung cancer and residential proximity to MTR sites over three decades. Additionally, we investigated the utility of LexisNexis data both for supplementing survey data and as an alternative to survey data.

2. Methods

This study was approved by the Medical Institutional Review Board at the University of Kentucky.

2.1 Data sources

Outcome and covariate data for this analysis originated from an age and gender frequency matched case-control study of lung cancer that examined exposure to trace elements in eastern Kentucky from 2012 through 2014, as reported previously by Urine and colleagues.⁵ That study collected and examined biological samples (toe nails) from lung cancer cases and controls to compare concentrations of trace elements known to be lung carcinogens. Further details are available in the article reporting the main findings of that study, which did not find higher trace elements concentrations in the toe nails of lung cancer cases relative to controls.⁵ Other information available from this case-control study, however, includes previous addresses for all cases and controls, which were gathered by in-person interviewers. We used these residential histories to examine residential proximity to MTR mining both alone and in combination with address data purchased later from LN to examine exposure over a longer time period than afforded by toe nail analysis. We obtained up to 10 addresses for each participant from LN. The address data LN compiled derives from a proprietary algorithm that draws from multiple data sources, including financial documents and government records. Dates of residence in these data include the month and year—or, often only the year—that an individual was first or last documented at a specific address. We therefore limited time precision in our analysis to years.

Several GIS data layers containing polygons representing the location and areal extent of individual MTR mining sites in the Appalachian region, for each year from 1985 through 2015, comprised the environmental exposure data for this study. SkyTruth.org, a non-profit that tracks resource extraction through remote sensing, compiled this data set by using Landsat imagery to discern changes to topography from year to year, as described

previously.¹² We eliminated small polygons (<0.5 km²) from the data set to limit analysis to mining sites most likely to produce a substantial amount of dust in a given year.

2.2 Residential histories

We assessed each source of addresses—LN and participant-provided addresses (PPA) from the previous case-control study—separately for duplicate records. For residences with overlapping time periods, we compared addresses by the similarity of the provided geographical coordinates (i.e., latitude and longitude), if present, specific numeric and textual elements, and Soundex scores. Soundex scores are generated from an algorithm that compares the phonetic properties of letter combinations.¹³ Additional standardization of addresses in the PPA data was necessary in order to ensure robust geocoding results.

To produce an address data set that combined information from both LN and PPA, we assessed duplication across data sets, using similar methods that were earlier applied to each individual data set. In instances where the data sets listed the same address, but with overlapping time periods, we created a time period for the combined address that spanned the earliest and latest dates from either data set. In instances where there were two addresses tied to the same year(s), we retained both to recognize that some individuals may split their time between two residences. In this manner we merged the two data sets to form a third, combined residential history data set for each case and control participant, using as much information from each data set as possible. Each data set was subsequently used in GIS models to create LN, PPA, and combined measures of exposure for each participant.

The algorithm designed by Westat Inc.,¹⁰ did not allow for overlap in residential history, or for the residence to be duplicated in a later time period, as we did. The algorithm employed by Hurley and colleagues, however, allowed for multiple addresses with the same move-in date, although last date of residence was assigned by choosing a date prior to the move-in date of the next residence. In this study, we were concerned that ignoring overlapping time periods would provide an inaccurate assessment of exposure, since a relatively high proportion (approximately 9%) of those employed here tend to work outside of the county in which they live, and commute times are high for a rural area; this indicates economic integration with nearby metropolitan areas.¹⁴ It may be relatively common for Central Appalachian residents to live in other areas for work, but return to family homes on weekends, or seasonally.

Because we wanted to compare the relative quality of address data, rather than geocoding algorithms, we geocoded all addresses in LN and PPA data sets using the USA_LocalComposite.loc address locator available in the ArcGIS Business Analyst 2018, Version 10.6 (Esri Redlands, California), even though some coordinates were already present in the LN data. Later, to compare the address quality of LN and PPA data sets, we tabulated the coordinate quality codes resulting from the geocoding process. Such codes indicate whether each set of residential coordinates was derived from (a) a full street address with house or building number, (b) an address range on a street segment or block, (c) the midpoint of a street (street name only), (d) the centroid of the ZIP code, (e) the centroid of a city, or (f) that the given address data could not be matched to a known location.

2.3 Exposure assessment

After plotting residential coordinates for all case and control residences in the GIS, as well as the MTR mining polygons for 1985 through 2015, we then created a model to implement a series of spatial joins to identify the size, in square kilometers, of any mines within one kilometer of each residence. We subsequently summed these values across years of occupancy at each residence, and then summed values across all residences for each individual. In this manner, we used the square kilometer area of the mine as a proxy for mining intensity, since larger MTR mines would presumably release larger amounts of dust, in a cumulative exposure metric. The one-kilometer distance was selected after also considering a 0.5-kilometer distance, but the latter was rejected due to a very small number of participants living in such close proximity to an MTR site. Lastly, we characterized all individuals' MTR exposure as none, below median, or above median. We completed this exposure assessment procedure using the LN data set alone, the PPA data set alone, and then the combined data set, producing three different exposure metrics for each individual. Regardless of the metric, the great majority of participants (LN, 89.4%; PPA, 86.7%, combined, 82.5%) had exposures that were zero, indicating they had never lived within 1 km of an MTR mine of any size. Splitting the remaining participants with non-zero exposure into above- and below-median groups allowed us to maintain the ability to look at multiple categories of exposure, while maximizing the number of participants in each group.

2.4 Statistical analysis

We characterized cases and controls using counts and percentages for age (<50, 50–64, 65–74, 75+), gender (male, female), BMI (underweight, <18.5; normal, 18.5–24.9; overweight, 25.0–29.9; and obese, 30.0+), race (white/other), education (less than high school, high school, some college, or college or more), household income (<\$20,000/year, \$20,000–49,999/year, \$50,000+/year, and unknown), pack years of cigarette smoking (none, <20, 20–39, 40+, and unknown), and each MTR exposure metric (none, below median, above median), and used chi-squared tests to examine their relationship to the lung cancer outcome. Due to small sample size there were some cells with expected counts less than five; in those instances we used Fisher's exact tests.

Covariates selected for three final multivariate logistic regression models included age and gender, as well as those that were significantly associated with the outcome, namely education, income, BMI, and pack-years of smoking. We retained age and gender in the final models because they are known to be associated with lung cancer risk. Additionally, as this was an age and gender frequency matched case control study, these were included to further adjust for any residual bias. We retained both education and income in the final models because they were not significantly associated among participants in this sample. Each model additionally included one of the three categorical (none, below median, above median) exposure metrics—derived from LN only, PPA only, or the combined residential histories—to compare the odds of MTR exposure over time among cases and controls, after adjustment for the potential covariates. We considered analyzing these measures of exposure as continuous variables, but subsequent analysis showed categorical exposures produced better fit in the final models.

To assess for a potential relationship between smoking and high MTR exposure that might lead to confounding in the final regression analysis, we cross-tabulated pack-years of smoking (0, 1–39, 40+) and level of MTR exposure (none/below median, above median) for cases and controls separately. Due to small cell counts, those with no MTR exposure were combined with those who had below median exposure and compared to those with above median exposure; additionally, we used Fisher's exact tests to assess for potential associations. We also examined the influence of pack-years of smoking in the final model by omitting it and noting any changes to ORs for other independent variables, and by adding an interaction between pack-years of smoking and MTR exposure.

We conducted all statistical analyses in Stata 15 (StataCorp, College Station, TX), and considered p-values less than 0.05 statistically significant.

3. Results

There were a total of 1041 addresses in the PPA data (median 2, IQR 1–2, range 1–16 per person), and 1364 addresses reported by LexisNexis (median 2, IQR 1–3, range 1–11 per person), for the 150 cases and 370 controls in the original study. The combined dataset contained 2047 addresses among all participants, with a median of 3 (IQR 2–5, range 1–23) addresses per person. Address quality for each source is summarized in Table 1. The LN data set was of considerably higher quality, as many PPA addresses included only the city of residence, rather than a full street address.

Characteristics of participants in relation to case-control status is displayed in Table 2. There were no significant differences between cases and controls with regards to age, gender, or race, although there were very few non-white participants. A larger proportion of controls were obese (42.2%) and overweight (38.2%), compared to cases (obese, 30.7%; overweight, 30.0%). Controls were also more highly educated, and a greater proportion had high incomes. More controls also reported zero pack-years of smoking (45.7% vs 2.7%). In general, the cases and controls were similar by sex, age and race, but cases were more likely to be of normal weight, generally had less education and income, and were less likely to smoke, compared to controls.

Table 3 displays the relationships between each of the MTR exposure metrics and case-control status. For each MTR exposure metric, cases consistently had more individuals with above-median exposure; however, they also had fewer in the below-median group than the controls in all instances. Still, none of the chi-squared tests suggested statistically significant differences in exposure between cases and controls in this unadjusted analysis.

We did not observe any statistically significant relationships between MTR exposure and smoking pack-years among cases or controls in any of the data sets (Table 4). In most cases, the proportion of cases and controls who smoked 40+ pack-years was actually slightly higher among those in the none / below median MTR exposure group.

The final logistic regression models, in Table 4, display adjusted odds ratios (ORs) for each exposure metric, and thus enable a comparison of their performance in estimating the relationship with lung cancer status. The greatest differences among the models appear in

ORs for these exposure metrics. In the combined model, cases had higher odds of high exposure to MTR mining (OR=1.5, $p=0.08$), and lower odds (OR=0.42, $p=0.05$) of low exposure compared to controls, but the results were marginally significant. The model with the LN-derived exposure metric produced similar results. The model with the PPA-derived exposure metric did not demonstrate a significant relationship, but the results were quite similar overall to the other models.

Besides these differences in OR estimates for the exposure metrics, there were only slight differences among models with regard to the measures of association between individual covariates and lung cancer status. In all three models, the age groups had similar ORs, except for the 75 and older group, where ORs were somewhat lower when using the PPA-derived exposure metric. Notably, obese BMIs were protective to a significant degree in the PPA model (OR=0.4, $p<0.01$) compared to a normal BMI.

When pack-years of smoking was omitted from the final model, there were only minor differences in ORs for MTR exposure (analysis not shown), leading to no difference in interpretation. To further assess this relationship we added an interaction between pack-years and dichotomized MTR exposure (i.e., none / below median, above median) to the final model; the results, however, were uninterpretable due to small cell sizes.

4. Discussion

This study assessed the use of LN data to augment survey data from a previously conducted case-control study of lung cancer in eastern Kentucky, specifically within the Central Appalachian region, an area with a legacy of resource extraction and elevated cancer rates.^{15,16} We found that using the LN residential history to supplement information from participant-provided addresses, which were generally of lower quality, allowed us to detect marginally significant relationship between lung cancer and previous residential proximity to MTR mining.

All three final models showed a similar trend with regard to the exposure of interest—specifically that odds of highest exposure were greater among cases—but the PPA-derived exposure measure did not show a significant effect, and other measures suggested marginally significant associations. We observed similar relationships between all covariates and lung cancer status in all models, with only slight discrepancies in demographic characteristics. This indicates that use of LN-derived residential histories did provide additional information for the final analysis that altered the results enough to reveal marginally significant associations with lung cancer risk in a small population-based sample of cases and frequency-matched controls. Although these results are of marginal statistical significance, they are nonetheless noteworthy because they were found using data from the same participants in an earlier study focused on more recent exposure that found no relationship to lung cancer.

Using LN address data provided notable advantages in this study. We were able to include residential histories for seven participants for whom we had no address information in the PPA data set. In many cases, the PPA data were so imprecise as to only confirm that an

individual lived outside of the exposure area (i.e., cities or states that were far removed from MTR sites). When PPA addresses were within the exposure area, the data were often not sufficiently precise for determining proximity (1 km or less) to MTR mining sites. The LN addresses generally provided more precise information in regards to residence, but often lacked residences found in the PPA data. Still, an overall lack of precision in address data was an important factor in our decision not to use inverse-distance weighting of MTR mine size when calculating the exposure metric. Unlike Jaquez et al.,⁶ who used survey data as a standard by which to gauge the quality and completeness of LN data, we found that our survey data were lacking the detail of LN addresses. This could be due in part, however, to our use of up to ten previous addresses per individual, rather than only three.

We did not anticipate observing the lowest ORs among participants with below median exposure, rather than none. This pattern, although of marginal significance at best, could indicate the effects of other unknown confounders or risk factors common to areas without any MTR mining that are not incorporated into our analysis. It seems likely that an individual's occupational history, for example, could influence their risk for lung cancer as well as their choice of residence, in terms of both location and housing quality. Both of these factors can influence exposure to airborne particulates associated with chronic respiratory diseases or lung cancer,^{17–23} but we were not able to include those covariates in this analysis.

This study and previous studies have demonstrated that use of address data from LN can enhance the completeness and quality of residential history survey data, but with limitations. Specifically, LN data may not accurately characterize dates of residence. While LN provides both first and last date that an individual was associated with an address, these data are drawn from several types of financial and administrative records, and may not accurately reflect actual dates of residence. Jaquez and colleagues found that years of residence may be inaccurate compared to participant provided information.⁶ Hurley et al., however, found that the LN data had high accuracy both spatially and temporally.⁷

Unlike previous studies employing this methodology, we did not have a high-quality standard by which to gauge the completeness and quality of LN data. There were some obvious ways the LN data were clearly superior in quality. Specifically, the street address field in the PPA data often contained narrative directions based on local landmarks or other information that was not helpful for geocoding. This did not occur in the LN data. Another potential difference with previous studies is that we chose to recognize as legitimate both (1) addresses during overlapping time periods, and (2) a repeated address during a later time period. The Westat Inc. algorithm did not retain such addresses in the final residential history, and Jaquez et al. did not describe their algorithms with sufficient detail to enable comparison.⁶ Hurley et al. recognized addresses that had the same move-in day, but move-out days were assigned based on the move-in date at the following address.⁷

Hurley and colleagues have suggested that LN data might capture fewer residences for African-Americans.⁷ A major limitation of this study was our small homogenous sample. Thus, we could not assess disparities related to race or ethnicity. Potential bias in LN related to race/ethnicity merits further study in other regions, but was simply not possible in this

sample from Central Appalachia, which is overwhelmingly white and non-Hispanic. Other limitations of our analysis include use of a proxy measure of exposure based on residence within one kilometer of an MTR site. There could be great variation in exposure to airborne dust and other particulates from MTR sites, depending not only on distance, but also on altitude, wind and weather patterns, and other factors not considered here. In this sense, our proxy measure of exposure might not be highly accurate. Additionally, allowing multiple addresses during the same time period in participants' residential histories could have overestimated their exposure. If such overestimation occurred more frequently among cases or controls, it could have biased our results.

5. Conclusion

Latency poses a known challenge when assessing environmental exposures in epidemiologic studies of cancer. Including commercially available data presents a solution to some of the difficulties in obtaining personal residential histories, but few studies have explored its utility in rural or isolated regions. We believe the residential histories we created recognize patterns of residential mobility that are relatively common in Appalachia, due its history and economic geography, but are perhaps rarer elsewhere in the US. It is likely that other uncommon or unique patterns of residential mobility are characteristic of, or more prevalent in, other regions. Future research should describe and account for such variation to realize the full potential of using LN data for exposure assessment in epidemiologic studies.

Acknowledgements

This research was supported by the University of Kentucky Center for Appalachian Research in Environmental Sciences through grant P30ES026529. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS. This research was also supported by the Biostatistics and Bioinformatics Shared Resource Facility of the University of Kentucky Markey Cancer Center through grant P30CA177558.

6. References

1. Hendryx M, O'Donnell K, Horn K. Lung cancer mortality is elevated in coal-mining areas of Appalachia. *Lung cancer* (Amsterdam, Netherlands). 2008;62(1):1–7.
2. Hendryx M Mortality from heart, respiratory, and kidney disease in coal mining areas of Appalachia. *International archives of occupational and environmental health*. 2009;82(2):243–249. [PubMed: 18461350]
3. Ahern MM, Hendryx M, Conley J, Fedorko E, Ducatman A, Zullig KJ. The association between mountaintop mining and birth defects among live births in central Appalachia, 1996–2003. *Environmental research*. 2011;111(6):838–846. [PubMed: 21689813]
4. Mountaintop Mining Overview. *Congressional Digest*. 2010;89(5):130.
5. Unrine JM, Slone SA, Sanderson W, et al. A case-control study of trace-element status and lung cancer in Appalachian Kentucky. *PLOS ONE*. 2019;14(2):e0212340. [PubMed: 30811496]
6. Jacquez GM, Slotnick MJ, Meliker JR, AvRuskin G, Copeland G, Nriagu J. Accuracy of commercially available residential histories for epidemiologic studies. *Am J Epidemiol*. 2011;173(2):236–243. [PubMed: 21084554]
7. Hurley S, Hertz A, Nelson DO, et al. Tracing a Path to the Past: Exploring the Use of Commercial Credit Reporting Data to Construct Residential Histories for Epidemiologic Studies of Environmental Exposures. *American Journal of Epidemiology*. 2017;185(3):238–246. [PubMed: 28073765]
8. Accident, Illness, and Injury and Employment Self-Extracting Files (Part 50 data). In: (US-MSHA) USDOL-MSaHA, ed2018.

9. Jones RR, DellaValle CT, Flory AR, et al. Accuracy of residential geocoding in the Agricultural Health Study. *International journal of health geographics*. 2014;13:37. [PubMed: 25292160]
10. Stinchcomb D, Roeser A. NCI/SEER Residential History Project Technical Report. Rockville, MD: Westat, Inc.;2016.
11. Wheeler DC, Wang A. Assessment of Residential History Generation Using a Public-Record Database. *International journal of environmental research and public health*. 2015;12(9):11670–11682. [PubMed: 26393626]
12. Pericak AA, Thomas CJ, Kroodsma DA, et al. Mapping the yearly extent of surface coal mining in Central Appalachia using Landsat and Google Earth Engine. *PLOS ONE*. 2018;13(7):e0197758. [PubMed: 30044790]
13. Drummond WJ. Address Matching: GIS Technology for Mapping Human Activity Patterns. *Journal of the American Planning Association*. 1995;61(2):240–251.
14. Mather M. Housing and commuting patterns in Appalachia. Citeseer; 2004.
15. Wilson RJ, Ryerson AB, Singh SD, King JBJCE, Biomarkers P. Cancer incidence in Appalachia, 2004–2011. 2016;25(2):250–258.
16. Moore TG Jr. Historical geography of economic development in Appalachian Kentucky, 1800–1930. 1984.
17. Brenner DR, McLaughlin JR, Hung RJ. Previous lung diseases and lung cancer risk: a systematic review and meta-analysis. *PLoS One*. 2011;6(3):e17479. [PubMed: 21483846]
18. Qu YL, Liu J, Zhang LX, et al. Asthma and the risk of lung cancer: a meta-analysis. *Oncotarget*. 2017;8(7):11614–11620. [PubMed: 28086224]
19. Tomczak A, Miller AB, Weichenthal SA, et al. Long-term exposure to fine particulate matter air pollution and the risk of lung cancer among participants of the Canadian National Breast Screening Study. *International journal of cancer*. 2016;139(9):1958–1966. [PubMed: 27380650]
20. Howden-Chapman P, Matheson A, Crane J, et al. Effect of insulating existing houses on health inequality: cluster randomised study in the community. *BMJ (Clinical research ed)*. 2007;334(7591):460.
21. Keall MD, Crane J, Baker MG, Wickens K, Howden-Chapman P, Cunningham M. A measure for quantifying the impact of housing quality on respiratory health: a cross-sectional study. *Environmental Health*. 2012;11(1):33. [PubMed: 22583775]
22. Pratt GC, Vadali ML, Kvale DL, Ellickson KM. Traffic, air pollution, minority and socio-economic status: addressing inequities in exposure and risk. *International journal of environmental research and public health*. 2015;12(5):5355–5372. [PubMed: 25996888]
23. Bell ML, Ebisu K. Environmental inequality in exposures to airborne particulate matter components in the United States. *Environ Health Perspect*. 2012;120(12):1699–1704. [PubMed: 22889745]

Table 1:

Geographic coordinate quality for each source of addresses

	<u>LN</u>		<u>PPA</u>		<u>Combined</u>	
	N	%	N	%	N	%
Full street address	533	39.1	158	15.2	590	28.8
Street address range	426	31.2	142	13.6	463	22.6
Street name (midpoint)	103	7.6	204	19.6	264	12.9
ZIP code centroid	300	22.0	112	10.8	345	16.9
City centroid	2	0.2	393	37.8	385	18.8
Unmatchable	0	0	32	3.1	0	0
Total addresses	1364	100	1041	100	2047	

Percentages might not sum to 100 due to rounding.

Table 2.

Demographic and personal characteristics of cases and controls

	Cases		Controls		Total		P-value ^a	
	N	%	N	%	N	%		
<u>Age</u>								
<50	23	15.3	43	11.6	66	12.7	0.66	
50–64	63	42.0	164	44.3	227	43.7		
65–74	45	30.0	120	32.4	165	31.7		
75+	19	12.7	43	11.6	62	11.9		
<u>Gender</u>								
Female	88	58.7	194	52.4	282	54.2	0.20	
Male	62	41.3	176	47.6	238	45.8		
<u>BMI</u>								
Underweight	8	5.3	5	1.4	13	2.5	<0.01	
Normal	51	34.0	67	18.3	118	22.8		
Overweight	45	30.0	140	38.2	185	35.8		
Obese	46	30.7	155	42.2	201	38.9		
<u>Race</u>								
Other	3	2.0	5	1.4	8	1.5	0.59	
White	147	98.0	365	98.7	512	98.5		
<u>Education</u>								
<HS	62	41.6	56	15.2	118	22.8	<0.01	
HS	48	32.2	138	37.4	186	35.9		
Some college	28	8.8	93	25.2	121	23.4		
College+	11	7.4	82	22.2	93	18.0		
<u>Income</u>								
<\$20k	9	52.7	82	22.2	161	31.0	<0.01	
\$20k–\$49,999k	36	24.0	126	34.1	162	31.2		
\$50k–\$99,999k	12	8.0	110	29.7	122	23.5		
Unknown	23	15.3	52	14.1	75	14.4		
<u>Pack-years</u>								
None	4	2.7	169	45.7	173	33.3	<0.01	
<20	21	14.0	57	15.4	78	15.0		
20–39	33	22.0	56	15.1	89	17.1		
40+	90	60.0	69	18.7	159	30.6		
Unknown	2	1.3	19	5.1	21	4.0		

^a Chi-square, or Fisher's exact test where expected cells counts were <5.

Table 3.Exposure^a to mountain top removal among cases and controls in eastern Kentucky

	Cases		Controls		P-value
	N	%	N	%	
<u>Combined-based exposure^a</u>					
Above median (>11.62)	16	10.7	31	8.4	0.21
Below median (11.62)	8	5.3	36	9.7	
None (0)	126	84.0	303	81.9	
<u>PPA-based exposure^a</u>					
Above median (>13.23)	12	8.2	22	6.0	0.25
Below median (13.23)	6	4.1	28	7.7	
None (0)	129	87.8	316	86.3	
<u>LN-based exposure^a</u>					
Above median (>6.0)	11	7.3	23	6.2	0.31
Below median (6.0)	3	2.0	18	4.9	
None (0)	136	90.7	329	88.9	

PPA, participant-provided addresses; LN, LexisNexis addresses; Combined, PPA and LN addresses combined and unduplicated.

^aExposure was defined as total area (km²) of all mines within 1 km of each participant's residence(s) during the study period. It was then further categorized into Above median, Below median, or None.

Table 4.

Cross-tabulation of smoking pack-years and MTR exposure among cases and controls by data source

	0 pack-years		1–39 pack-years		40+ pack-years		Total	
LN (controls)	n	%	n	%	n	%	n	%
None / Below Median	158	48.02	104	31.61	67	20.36	329	100.00
Above Median	11	50.00	9	40.91	2	9.09	22	100.00
Total	169	48.15	113	32.19	69	19.66	351	100.00
Fischer's Exact p-value	0.40							
LN (cases)	n	%	n	%	n	%	n	%
None / Below Median	4	2.92	49	35.77	84	61.31	137	100.00
Above Median	0	0.00	5	45.45	6	54.55	11	100.00
Total	4	2.70	54	36.49	90	60.81	148	100.00
Fischer's Exact p-value	0.82							
PPA (controls)	n	%	n	%	n	%	n	%
None / Below Median	158	48.32	103	31.50	66	20.18	327	100.00
Above Median	9	42.86	10	47.62	2	9.52	21	100.00
Total	167	47.99	113	32.47	68	19.54	348	100.00
Fischer's Exact p-value	0.28							
PPA (cases)	n	%	n	%	n	%	n	%
None / Below Median	3	2.24	50	37.31	81	60.45	134	100.00
Above Median	1	8.33	3	25.00	8	66.67	12	100.00
Total	4	2.74	53	36.30	89	60.96	146	100.00
Fischer's Exact p-value	0.27							
Combined (controls)	n	%	n	%	n	%	n	%
None / Below Median	155	48.29	100	31.15	66	20.56	321	100.00
Above Median	14	46.67	13	43.33	3	10.00	30	100.00
Total	169	48.15	113	32.19	69	19.66	351	100.00
Fischer's Exact p-value	0.26							
Combined (cases)	n	%	n	%	n	%	n	%
None / Below Median	3	2.27	48	36.36	81	61.36	132	100.00
Above Median	1	6.25	6	37.50	9	56.25	16	100.00
Total	4	2.70	54	36.49	90	60.81	148	100.00
Fischer's Exact p-value	0.48							

Table 5.

Adjusted logistic regression for case-control study of lung cancer and mountaintop removal mining, by address data source used for exposure assessment

	PPA model				LN model				Combined model			
	OR	95% CI	P-value	OR	95% CI	P-value	OR	95% CI	P-value	OR	95% CI	P-value
<u>Age</u>												
75+	1.08	0.40	2.94	0.34	1.42	0.53	3.79	0.15	1.32	0.50	3.52	0.15
65–74	0.63	0.29	1.35	0.25	0.70	0.33	1.49	0.18	0.61	0.28	1.31	0.11
50–64	0.61	0.29	1.30	0.19	0.69	0.32	1.49	0.20	0.65	0.31	1.38	0.19
<50	Ref.				Ref.				Ref.			
<u>Gender</u>												
Male	0.39	0.23	0.65	<0.01	0.42	0.25	0.70	<0.01	0.41	0.24	0.69	<0.01
Female	Ref.				Ref.				Ref.			
<u>BMI</u>												
Obese	0.41	0.22	0.78	0.01	0.40	0.21	0.75	0.01	0.39	0.20	0.73	0.01
Overweight	0.48	0.25	0.92	0.04	0.48	0.25	0.91	0.06	0.46	0.24	0.88	0.05
Underweight	2.29	0.53	9.96	0.06	1.83	0.44	7.54	0.09	1.85	0.44	7.75	0.09
Normal	Ref.				Ref.				Ref.			
<u>Income</u>												
Unknown	2.42	0.92	6.36	0.20	2.86	1.11	7.39	0.11	2.64	1.03	6.80	0.14
Low	2.47	1.05	5.80	0.08	2.75	1.18	6.41	0.06	2.56	1.10	5.96	0.08
Middle	1.38	0.60	3.16	0.33	1.45	0.64	3.29	0.26	1.43	0.63	3.25	0.32
High	Ref.				Ref.				Ref.			
<u>Education</u>												
<HS	2.12	0.80	5.58	0.04	2.00	0.78	5.11	0.03	2.07	0.81	5.31	0.03
HS	0.98	0.39	2.46	0.17	0.89	0.36	2.17	0.12	0.88	0.36	2.17	0.10
Some College	1.39	0.54	3.60	0.79	1.24	0.49	3.15	0.93	1.29	0.51	3.28	0.86
College +	Ref.				Ref.				Ref.			
<u>Pack-years</u>												
40+	68.61	21.46	219.30	<0.01	71.72	22.27	230.98	<0.01	71.70	22.36	229.86	<0.01
20–39	24.28	7.47	78.90	<0.01	29.39	9.02	95.80	<0.01	27.41	8.49	88.47	<0.01
<20	18.52	5.57	61.60	0.05	18.33	5.53	60.78	0.13	17.65	5.36	58.17	0.13
Unknown	3.44	0.56	21.16	0.09	5.13	0.80	32.78	0.21	4.77	0.75	30.30	0.19
None	Ref.				Ref.				Ref.			
<u>Exposure^a</u>												
Above Median	1.77	0.71	4.41	0.12	1.71	0.66	4.40	0.05	1.81	0.80	4.11	0.03
Below Median	0.61	0.20	1.86	0.19	0.28	0.07	1.19	0.04	0.45	0.17	1.18	0.03
None	Ref.				Ref.				Ref.			

^aExposure was defined as total area (km²) of all mines within 1 km of each participant's residence(s) during the study period. It was then further categorized into Above median, Below median, or None.