A Section 508–conformant HTML version of this article is available at https://doi.org/10.1289/EHP4913.

Constituents of Household Air Pollution and Risk of Lung Cancer among Never-Smoking Women in Xuanwei and Fuyuan, China

Roel Vermeulen,[1] George S. Downward,[1*] Jinming Zhang,[2*] Wei Hu,[2] Lützen Portengen,[1] Bryan A. Bassig,[2] S. Katharine Hammond,[3] Jason Y.Y. Wong,[2] Jihua Li,[4] Boris Reiss,[5] Jun He,[4] Linwei Tian,[6] Kaiyun Yang,[7] Wei Jie Seow,[8,9] Jun Xu,[10] Kim Anderson,[11] Bu-Tian Ji,[2] Debra Silverman,[2] Stephen Chanock,[2] Yunchao Huang,[7†] Nathaniel Rothman,[2†] and Qing Lan[2†]*

[1]Institute for Risk Assessment Sciences, Division of Environmental Epidemiology, Utrecht University, Utrecht, Netherlands
[2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, Maryland, USA
[3]Environmental Health Sciences Division, School of Public Health, University of California, Berkeley, California, USA
[4]Qujing Center for Diseases Control and Prevention, Qujing, Yunnan, China
[5]Department of Community, Environment & Policy, Mel & Enid Zuckerman College of Public Health, University of Arizona, Arizona, USA
[6]Division of Epidemiology and Biostatistics, School of Public Health, University of Hong Kong, Hong Kong, China
[7]Department of Thoracic Surgery, Third Affiliated Hospital of Kunming Medical University (Yunnan Cancer Hospital, Yunnan Cancer Center), Kunming, Yunnan, China
[8]Saw Swee Hock School of Public Health, National University of Singapore, Singapore
[9]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore
[10]School of Public Health, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong, China
[11]Department of Environmental and Molecular Toxicology, College of Agricultural Sciences, Oregon State University, Corvallis, Oregon, USA

Background: Lung cancer rates among never-smoking women in Xuanwei and Fuyuan in China are among the highest in the world and have been attributed to the domestic use of smoky (bituminous) coal for heating and cooking. However, the key components of coal that drive lung cancer risk have not been identified.

Objectives: We aimed to investigate the relationship between lifelong exposure to the constituents of smoky coal (and other fuel types) and lung cancer.

Methods: Using a population-based case–control study of lung cancer among 1,015 never-smoking female cases and 485 controls, we examined the association between exposure to 43 household air pollutants and lung cancer. Pollutant predictions were derived from a comprehensive exposure assessment study, which included methylated polycyclic aromatic hydrocarbons (PAHs), which have never been directly evaluated in an epidemiological study of any cancer. Hierarchical clustering and penalized regression were applied in order to address high colinearity in exposure variables.

Results: The strongest association with lung cancer was for a cluster of 25 PAHs [odds ratio (OR): 2.21; 95% confidence interval (CI): 1.67, 2.87 per 1 standard deviation (SD) change], within which 5-methylchrysene (5-MC), a mutagenic and carcinogenic PAH, had the highest individual observed OR (5.42; 95% CI: 0.94, 27.5). A positive association with nitrogen dioxide (NO_2) was also observed (OR: 2.06; 95% CI: 1.19, 3.49). By contrast, neither benzo(*a*)pyrene (BaP) nor fine particulate matter with aerodynamic diameter ≤ 2.5 μm ($PM_{2.5}$) were associated with lung cancer in the multipollutant models.

Conclusions: To our knowledge, this is the first study to comprehensively evaluate the association between lung cancer and household air pollution (HAP) constituents estimated over the entire life course. Given the global ubiquity of coal use domestically for indoor cooking and heating and commercially for electric power generation, our study suggests that more extensive monitoring of coal combustion products, including methylated PAHs, may be warranted to more accurately assess health risks and develop prevention strategies from this exposure. https://doi.org/10.1289/EHP4913

Introduction

The incidence of lung cancer among female never smokers in Asia is among the highest in the world (Couraud et al. 2012; Epplein et al. 2005; Pelosof et al. 2017; Samet et al. 2009). There is a large body of evidence identifying household air pollution (HAP) as a key risk factor for lung cancer among never smokers. Approximately half of the world's population is still exposed to HAP from domestic cooking and/or heating with solid fuels (wood, coal, biomass) (Barone-Adesi et al. 2012; Lan et al. 2002; Sisti et al. 2012). The lung cancer rates in nonsmoking women in Xuanwei, China, are among the highest in the world (Barone-Adesi et al. 2012; Chapman et al. 1988; Mumford et al. 1987). Residents of Xuanwei live primarily in rural areas, working as subsistence farmers who use solid fuels for domestic heating and cooking. Previous epidemiological and experimental studies have shown that the cause of this excess cancer risk is the use of bituminous coal, locally referred to as smoky coal, for indoor cooking and heating (Barone-Adesi et al. 2012; Chapman et al. 1988; Ho et al. 2016; Large et al. 2009; Lui et al. 2017a, 2017b; Mumford et al. 1987, 1989; Tian 2005; Tian et al. 2008). When compared with alternative fuels typically available in the area [including a carboniferous anthracite (smokeless) coal], smoky coal use was associated with a 100-fold increase in lung cancer mortality among nonsmoking women (Barone-Adesi et al. 2012). The key drivers of this excess risk remain elusive however, because no study to date has been able to estimate specific exposures for individuals in this population over their lifetime and then directly link those exposure estimates to lung cancer risk.

Given the widespread global use of coal for domestic heating and cooking and for power generation, the striking lung cancer excess in this region of China provides an important scientific and public health opportunity to further our mechanistic understanding of the carcinogenicity of coal, more comprehensively

*These authors contributed equally to this work.

Address correspondence to Roel Vermeulen, PhD, Institute for Risk Assessment Sciences (IRAS), Utrecht University, P.O. Box 80178, 3508 TD, Utrecht, Netherlands. Telephone: +31 30 253 9448. Fax: +31 30 253 9494.

Supplemental Material is available online (https://doi.org/10.1289/EHP4913).

†These authors co-supervised this work.

The authors declare they have no actual or potential competing financial interests.

Received 18 December 2018; Revised 12 July 2019; Accepted 13 August 2019; Published 5 September 2019.

evaluate its health risks, and develop strategies for prevention. We therefore conducted a comprehensive large-scale population-based lung cancer case–control study and a complementary exposure assessment study in never-smoking women in the counties of Xuanwei and Fuyuan (Downward et al. 2014a, 2014b; Downward 2015; Downward et al. 2016, 2017; Hu et al. 2014; Seow et al. 2016) We evaluated an array of HAP constituents, including traditional markers of air pollution, such as particulate matter with aerodynamic diameter less ≤ 2.5 μm ($PM_{2.5}$), black carbon (BC), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and the 16 priority U.S. Environmental Protection Agency (U.S. EPA) polycyclic aromatic hydrocarbons (PAHs), including benzo-(*a*)pyrene (BaP) (Downward et al. 2014b, 2016, 2017; Hu et al. 2014; Seow et al. 2016). To more comprehensively assess risks from specific PAHs, we analyzed an additional 23 PAHs, including multiple methylated subspecies, and also measured radon, 23 trace elements (including arsenic), and crystalline silica (quartz) in air (Downward 2015; Downward et al. 2017).

Methods

Data Collection

Study design. We conducted a population-based case–control study of lung cancer among never-smoking women from Xuanwei and Fuyuan, China. The enrollment methods have been described previously (Wong et al. 2019). Cases were enrolled from the six hospitals that diagnosed the majority of lung cancer cases in Xuanwei and Fuyuan. Cases were defined as newly diagnosed lung cancers [*International Classification of Disease, Ninth Revision* (ICD-9; CDC 2013) code 162] among female never smokers between the years of 2006 and 2013. Cases were required to be aged 18–79 y old (at the time of diagnosis), currently living in Xuanwei or Fuyuan, have lived in these counties for at least 1 y, and have had no previous history of cancer diagnoses. Cases were initially identified on the basis of clinical presentation and radiological imaging and subsequently provided sputum samples for cytological analysis and tuberculosis (TB) testing. A subgroup of cases had more extensive diagnostic testing; however, cases without a confirmed histological or cytological diagnosis of lung cancer, clinical evidence, or radiologic findings (i.e., chest X-ray, computed tomography scan), and lack of evidence for other conditions (e.g., TB) were used as the basis for the diagnosis of lung cancer. Cases that were not confirmed based on diagnostic tests were followed up for vital status and cause of death for at least 3 y after enrollment and classified as high-probability lung cancer cases if they did not have another diagnosis explaining their initial presentation or cause of death. A total of 1,060 eligible lung cancer cases were considered, of which 706 were confirmed and 354 were high-probability cases.

We also enrolled a total of 498 population-based controls, frequency matched to cases by age (± 5 y). The case–control ratio was partly determined by financial and practical considerations (control interviews were much more difficult to arrange), while the total sample size was considered sufficient for statistical power based on the very strong associations between smoky coal use and lung cancer risk in this region (Barone-Adesi et al. 2012; Lan et al. 2008) Controls were never smokers who did not have a previous cancer diagnosis, were currently living in Xuanwei or Fuyuan, and had lived in these counties for >1 y.

Control sampling was based on four geographic levels from largest to smallest, as follows: *a*) commune, *b*) administrative villages ("dadui"), *c*) natural villages/settlements, and *d*) individuals. Information on the population density across age strata was available for all communes in Xuanwei and Fuyuan from the National Bureau of Statistics of China. Along with the predicted age distribution of cases in these provinces, this information was used to randomly select natural villages nested within administrative villages and communes from which controls should be sampled. Subsequently, the field team of interviewers visited each natural village, selected three potential participants fitting the eligibility criteria, and randomly recruited one of them.

Cases and controls were interviewed using a detailed questionnaire that collected information on residential history, fuel use, and established or suspected risk factors for lung cancer.

We excluded 41 cases and 13 population-based controls because they reported using fuels or stove types that were not included in the exposure survey that was used for exposure assessment (i.e., coal deposit or stove type used) and an additional 4 cases because of missing information on one of the questionnaire items that was used to adjust for potential confounding in the models ("having sufficient food before marriage"). Participation rates were 84% for cases and 89% for controls. The present investigation was based on data from 1,015 cases and 485 controls.

Ethics. All participants provided written informed consent prior to participating in the study. The study protocol was approved by the institutional review boards of the National Cancer Institute and China National Environmental Monitoring Center.

Questionnaire. Information on demographics, lifestyle factors (including exposure to secondhand smoke), socioeconomic status (SES) (including food sufficiency before marriage), medical history, and household characteristics (including lifetime history of fuel and stove use, type of solid fuels used, and the mine from which coal was sourced) was collected through an administered questionnaire.

Coal deposit assignment and estimation of household air pollutant exposures. The geographic location of coal deposits used by study subjects and associated lung cancer risk by deposit are shown in Figure S1. Coal deposits were classified as smoky coal (i.e., coking coal, one-third coking, meager lean, gas fat) or smokeless coal based on the Chinese state standard coal classification (Chen 2000). Coal deposits from which the household coal originated were assigned on the basis of self-reported fuel sources. Lung cancer risks based on the aggregation of coal deposits that shared common characteristics were previously analyzed and are shown in Figure S1 (Wong et al. 2019). If participants reported that their coal source was a local coal mine, deposits were assigned through Bayesian predictive modeling, where proximity and preference of inhabitants of the same village were the primary predictors.

The Bayesian predictive modeling used a multinomial choice model for each village, which included all mines within a 30-km radius of the village center (selected because this represented the upper 90% of all observed distances between subjects and their reported coal mine). Probabilities for each mine were derived based on the observed choices of other individuals within the same village and a linear effect of straight-line distance to the village center (selecting for mines closer to the participant's village). To this end, probabilities resulted in 50% lower odds of selection for each additional 10-km increase in distance from the village center. To take into account information on reported coal type while allowing reporting errors to occur (previously reported to be approximately 10%) (Downward et al. 2014a), mines that produced the coal type reported were assigned fivefold higher odds of selection.

Selection probabilities for each deposit layer were calculated by summing up selection probabilities for individual mines that were mining that layer. The current analyses used the most likely predicted deposit layer for each participant who reported using coal as their primary fuel source.

Exposure assessment was performed through determinant based modeling as previously described (Downward et al. 2014b,

2016; Hu et al. 2014; Seow et al. 2016). Briefly, we enrolled 163 households and their never-smoking female heads, selected from 30 villages throughout Xuanwei and Fuyuan, for an exposure survey in which multiple indoor and personal air samples were collected. Village and subject selection was targeted to represent the major geographical regions, solid fuels, and stove designs in use and to provide a population reflective of that within the case–control study.

Personal and indoor measurements were collected over two sequential 24-h periods in 2008 and 2009. Samples of $PM_{2.5}$, BC, quartz, trace elements, and particle-bound PAHs and methylated PAHs (mPAHs) were collected on 37-mm Teflon filters. Indoor measurements of NO_2 and SO_2 were collected using passively diffusing filters (Ogawa). Radon was evaluated through the deployment of 90-d passively collecting detectors. Approximately 50% of subjects were visited again in a second season to allow seasonal adjustment of findings. Because levels of crystalline silica, trace elements, and radon were not associated with either fuel use or the use of different coal deposits, they were not included in any subsequent analyses (Downward 2015).

An overview of the constituents measured, by stove and fuel type, is given in Tables S1 and S2. Predictive linear mixed-effect models were constructed for each pollutant, where villages and individual subjects were assigned as random effects (Downward 2015, 2016; Hu et al. 2014; Seow et al. 2016). Model construction was performed under a supervised stepwise procedure wherein variables were individually considered for model inclusion on the basis of goodness of fit and Akaike information criteria. Additional information is available in Text S1.

These predictive models were applied to the self-reported histories of stove and fuel use within the case–control data set to assign annual geometric mean predicted exposure levels. For case–control subjects who reported using more than one fuel/stove in a given year, predictions from each permutation were combined based on the proportion of time spent using each permutation (e.g., a person using smoky coal and wood in equal amounts would have their final assignment calculated from a 50% smoky coal and 50% wood prediction).

Statistical Analysis

We followed two approaches to investigate the relation between exposure to HAP and lung cancer case–control status. We first identified clusters of highly correlated HAP constituents using a hierarchical clustering method that is described in more detail in the next paragraph. Regression models were then fitted that related case–control status to cluster scores. For the second approach, we fitted models using the full set of HAP constituents ($n = 43$). All models were adjusted for age, self-reported exposure to secondhand smoke, and self-reported food sufficiency before marriage as an indicator of SES. Food sufficiency is the most appropriate metric of SES in this population, which comprises mostly rural subsistence farmers with nonguaranteed food security and relatively little contrast in education.

Cluster analysis. We performed a hierarchical cluster analysis on exposures of the controls after a Box Cox transformation to reduce skewness and improve symmetry of the distributions (Box and Cox 1964). The distance matrix was calculated using standard Euclidean distances, and the complete linkage method was used to determine the cluster sequence (Defays 1977). The number of clusters to extract was decided by maximizing the average silhouette width of the resulting clusters (Rousseeuw 1987). For clusters that consisted of a single exposure, a cluster score was calculated by assigning the (transformed) exposure level. For clusters that consisted of more than one exposure, the cluster score was calculated by assigning the first component score from a principal component analysis on all (transformed) exposures in that cluster. To evaluate whether clusters and their associations with cancer differed notably between the overall study population and those using smoky coal, cluster analysis was performed separately for the full study population and for those using smoky coal.

Effect estimation. We used regularized logistic regression to evaluate the association between estimated exposures and lung cancer case–control status. We used multipollutant models because of concerns about uncontrolled confounding by moderate or highly correlated coexposures in single-exposure models (Agier et al. 2016). Investigating health outcomes associated with correlated exposures using multipollutant models is challenging because of the sometimes low precision of coefficient estimates that results from the large number of potential multicollinear covariates and the potential for introducing rather than removing bias (Schisterman et al. 2017; Weisskopf et al. 2018). To address these challenges and improve interpretability of the results, we used a Bayesian regularized regression method to fit our multipollutant models. We preferred a Bayesian approach over a frequentist regularization method (like the well-known lasso or elastic net) because the full posterior distribution allows more complete inference and assessment of model uncertainty (Casella et al. 2010).

We used the horseshoe estimator as originally suggested by Carvalho et al. (2010) by formulating shrinkage priors on the regression coefficients in a Bayesian regression model. We followed recommendations for further modifications of the horseshoe prior and hyperparameter settings that were suggested by Piironen and Vehtari (2017) and implemented in the R package brms (Bürkner 2017) (R version 2.3.1; R Development Core Team). The hyperparameter for the scale of the global shrinkage parameter was selected assuming that all of the confounders, but no more than 50% of the exposures, were associated with lung cancer, while the degrees of freedom (df) for this parameter were left at the default value of 1. The scale parameter for the slab (i.e., the regularization prior for coefficient values of effective variables) was set to 2.5, the df for the slab to 4, and the df for the local shrinkage parameters to 1. We increased the adapt_delta parameter of the no-u-turn sampler (Betancourt 2018) to 0.99 to avoid divergent transitions. We checked the sensitivity of our results to these assumptions by comparing them with results from models that were more strongly regularized (i.e., under the assumption that only a single exposure was related to the outcome) but found that results were not very sensitive to the exact choice. An additional sensitivity analysis using models with priors that implied very little regularization [i.e., a student *t*-test prior with 7 df and a scale of 2.5 (Gelman et al. 2008)] resulted in point estimates that were broadly similar to the regularized models but had considerably wider confidence intervals (CIs) (Table S3).

We used 4 chains and 4,000 iterations per chain, discarding the first 2,000 iterations as burn-in. The model was checked for convergence using the Gelman-Rubin diagnostic (Brooks and Gelman 1998).

Receiver operating characteristic analysis. To assess model fit of the regularized regression models and to investigate to what extent the models based on estimated pollutant levels improved on models based on coal deposit information only (Figure S1), we calculated the area under the curve (AUC) for the prediction of lung cancer case status using a receiver operating characteristic (ROC) analysis (Fawcett 2006). Case–control probabilities were estimated using 10-fold cross-validation to avoid overfitting of the prediction model. Models were adjusted for the same covariates as in the main analysis above using the approach described by Yu et al. (2012), who suggest including estimated covariate effects as a fixed offset in the fold-specific prediction models.

Sensitivity analyses. Exposure contrast was partly due to differences between subjects who mainly used smokeless coal or wood and subjects who mainly used smoky coal. Although this may improve the precision of our regression estimates, it also complicates interpretation because of the potential for introducing aggregation bias and unmeasured confounding (Navidi et al. 1994). We therefore conducted sensitivity analyses that excluded participants who had used smokeless coal and/or wood for more than 2 y. A total of 859 cases (85%) and 182 controls (38%) had not used smokeless coal and/or wood for more than 2 y (i.e., the smoky coal subpopulation).

We also investigated the potential for a lagged effect between modeled exposures and outcome. However, the rank correlations between 10-y lagged cumulative exposure and unlagged cumulative exposure was >0.98 for all HAP constituents. Given this high correlation, no additional lag analysis was performed.

Results

Population Demographic and Exposure Characteristics

An overview of participant characteristics for cases and controls is provided in Table 1 with commune-specific populations in Table S4. A notable difference was that 96% of cases vs. 79% of controls had ever used smoky coal. Additionally, controls were more likely to have experienced food sufficiency than cases (62% vs. 57%, respectively). An overview of estimated exposure to all 43 retained HAPs by case–control status is provided in Figure S2. In general, cases experienced higher exposures of individual constituents than controls. A heatmap showing Spearman correlation coefficients between the different exposures is shown in Figure S3.

Hierarchical cluster analysis identified six exposure clusters based on data from all controls. An overview of the clusters, alongside pollutant abbreviations, is presented in Table 2. Following review of the clusters, they were designated as follows: PAH25 [consisting of 25 PAHs, including BaP and 5-methylchrysene (5-MC)], WB7 (consisting of seven wood burning–associated exposures, including $PM_{2.5}$ and retene), PAH7, PAH2, and two single-exposure clusters consisting of NO_2 and SO_2, respectively. Correlations between clusters ranged from −0.51 (PAH2:PAH7) to 0.64 (PAH25:NO_2) with a median coefficient of 0.38 (Table S5). Table S6 shows the correlations between the different clusters and between each individual exposure and each cluster. Within each cluster, the individual exposure metrics are naturally positively correlated with each other. Within the PAH25 cluster in particular, correlations were typically greater than 0.9 (Figure S3).

The cluster analysis for the subset of participants that used only or mainly smoky coal resulted in the identification of five clusters, which were designated PAH29+ (29 PAHs plus NO_2 and SO_2), PAH2, PAH5+, WB2, and a single-exposure cluster consisting of retene. Correlations between clusters ranged from −0.44 to 0.79 with a median coefficient of 0.48 (Table S7). Within each cluster, the individual exposure metrics again positively correlated with each other (Table S8).

Lung Cancer and Household Air Pollution Components

Results from the regularized regression analysis of the relationship between the above clusters and lung cancer are presented in Figure 1 for both the full study population and the smoky coal subpopulation (see Tables S9 and S10 for numerical values). In the full population, the PAH25 and NO_2 clusters were the most positively associated with lung cancer with odds ratios (ORs) of 2.21 (95% CI: 1.67, 2.87) and 2.27 (95% CI: 1.48, 3.55 for a 1 standard deviation increase in exposure, respectively). Exposures to pollutants in clusters WB7 and SO_2 were associated with a lower odds of lung cancer [ORs: 0.24 (95% CI: 0.19, 0.31) and 0.55 (95% CI: 0.43, 0.70), respectively], suggesting that these may reflect use of noncoal fuels or less toxic coals.

Results from the (regularized) model that included all individual exposures are presented in Figure 2 (see Table S3 for numerical values). The highest individual OR observed was for 5-MC within cluster PAH25 (OR: 5.42; 95% CI: 0.94, 37.5). Positive ORs were also estimated for other PAHs in this cluster, including benz(*a*)-anthracene (OR: 3.15; 95% CI: 0.65, 135), and benzo(*g,h,i*)perylene (OR: 1.84; 95% CI: 0.52,75). Similarly to the cluster analysis above, NO_2 was also positively associated with lung cancer (OR, 2.06; 95% CI: 1.19, 3.49). Exposure to BaP was not associated with lung cancer (OR: 0.99; 95% CI: 0.23, 4.00).

When examining the results of the regularized regression analysis upon the clusters identified within the smoky coal subpopulation (Figure 1B; numerical values available in Table S10), positive (albeit nonsignificant) ORs were observed for PAH29+

Table 1. Population characteristics of a case–control study of lung cancer among never-smoking women in Xuanwei and Fuyuan, China.

Characteristic	Cases ($n = 1{,}015$)	Controls ($n = 485$)	p-Value
Age (y) [mean (SD)]	54.9 (10.4)	54.8 (11.4)	0.92
County [n (%)[a]]			
Xuanwei	630 (62.1)	288 (59.4)	0.35
Fuyuan	385 (37.9)	197 (40.6)	—
Ever environmental tobacco smoke exposure [n (%)]			
No	35 (3.4)	9 (1.9)	0.12
Yes	980 (96.6)	476 (98.1)	—
Food sufficiency before marriage [n (%)]			
More than enough	106 (10.4)	30 (6.2)	0.02
Just enough	326 (32.1)	154 (31.8)	—
Not enough	583 (57.4)	301 (62.1)	—
Ever used smoky coal [n (%)]			
No	40 (3.9)	100 (20.6)	<0.0001
Yes	975 (96.1)	385 (79.4)	—
Used smokeless coal or wood [n (%)]			
Never	846 (83.3)	142 (29.3)	<0.0001
≤2 y across the lifetime	10 (1.0)	40 (8.2)	—
>2 y across the lifetime	159 (15.7)	303 (62.5)	—
Smoky coal use per year in lifetime (tons) [mean (SD)]	182 (100)	149 (122)	<0.0001

Note: Continuous variables were compared using Wilcoxon rank sum tests. Categorical variables were compared using chi-square or Fisher's exact tests. Minor discrepancy in some counts are due to missing data. —, no data; SD, standard deviation.

[a]The total eligible populations of Xuanwei and Fuyuan, when developing the present study, were 442,975 and 207,302, respectively. Commune-specific populations are presented in Table S4.

Table 2. Measured constituents and assigned clusters in the full population and smoky coal subgroup.

Pollutant	Full name	Full population cluster	Smoky population cluster
NO_2	Nitrogen dioxide	NO_2	PAH29 +
SO_2	Sulfur dioxide	SO_2	PAH29 +
ANT	Anthanthrene	WB7	WB2
BC	Black carbon	WB7	PAH5 +
CdP	Cyclopenta(*c,d*)pyrene	WB7	WB2
FLT	Fluoranthene	WB7	PAH29 +
$PM_{2.5}$	Fine particulate matter with aerodynamic diameter ≤2.5 μm	WB7	PAH29 +
PYR	Pyrene	WB7	PAH29 +
RET	Retene	WB7	RET
DlP	Dibenzo(*a,l*)pyrene	PAH2	PAH2
NkF	Naphtho(*2,3,k*)fluoranthene	PAH2	PAH2
6MC	6-Methylchrysene	PAH7	PAH29 +
BbP	Benzo(*b*)perylene	PAH7	PAH5 +
BjA	Benz(*j,e*)aceanthrylene	PAH7	PAH5 +
DhP	Dibenzo(*a,h*)pyrene	PAH7	PAH5 +
DiP	Dibenzo(*a,i*)pyrene	PAH7	PAH5 +
DMBA	7,12-Dimethylbenz(*a*)anthracene	PAH7	PAH29 +
NaP	Naphtho(*2,3,a*)pyrene	PAH7	PAH5 +
1MP	1-Methylpyrene	PAH25	PAH29 +
5MC	5-Methyl chrysene	PAH25	PAH29 +
BaA	Benz(*a*)anthracene	PAH25	PAH29 +
BaC	Benzo(*a*)chrysene	PAH25	PAH29 +
BaF	Benzo(*a*)fluorene	PAH25	PAH29 +
BaP	Benzo(*a*)pyrene	PAH25	PAH29 +
BbF	Benzo(*b*)fluorene	PAH25	PAH29 +
BbT	Benzo(*b*)fluoranthene	PAH25	PAH29 +
BcF	Benzo(*c*)fluorene	PAH25	PAH29 +
BeP	Benzo(*e*)pyrene	PAH25	PAH29 +
BgP	Benzo(*g,h,i*)perylene	PAH25	PAH29 +
BjF	Benzo(*j*)fluoranthene	PAH25	PAH29 +
BkF	Benzo(*k*)fluoranthene	PAH25	PAH29 +
CHR	Chrysene	PAH25	PAH29 +
COR	Coronene	PAH25	PAH29 +
DBA	Dibenzo(*a,h*)anthracene	PAH25	PAH29 +
DBF	Dibenzo(*a,e*)fluoranthene	PAH25	PAH29 +
DBT	Dibenzothiophene	PAH25	PAH29 +
DelP	Dibenzo(*e,l*)pyrene	PAH25	PAH29 +
DeP	Dibenzo(*a,e*)pyrene	PAH25	PAH29 +
IPY	Indeno(*1,2,3*)pyrene	PAH25	PAH29 +
NbF	Naphtho(*1,2,b*)fluoranthene	PAH25	PAH29 +
NeP	Naphtho(*2,3,e*)pyrene	PAH25	PAH29 +
NjF	Naphtho(*2,3,j*)fluoranthene	PAH25	PAH29 +

Note: PAH, polycyclic aromatic hydrocarbon.

(OR: 1.13; 95% CI: 0.97, 1.54) and PAH2 (OR: 1.11; 95% CI: 0.98, 1.51). When examining the estimated ORs and 95% CIs from the full multipollutant model for the smoky coal subpopulation (Figure 3; numerical values available in Table S11), 5-MC again had the strongest individual OR for lung cancer (OR: 28.5; 95% CI: 1.02, 772), followed by benzo(*g,h,i*)perylene (OR: 9.74; 95% CI: 0.34, >4,000) and coronene (OR: 4.64; 95% CI: 0.39, 814). A highly protective OR (<0.01) was observed for retene, a PAH primarily associated with wood combustion (i.e., a fuel not associated with lung cancer in Xuanwei). Similarly to the total population, increased ORs were observed for NO_2 (OR: 2.21; 95% CI: 0.96, 5.71).

A graphical comparison of the coefficients from the (regularized) regression models fitted to the full population data or data from the smoky coal subpopulation is provided in Figure S4. Overall, there was high agreement between both models (Pearson's correlation of 0.8), and the positive associations observed in the full population typically remained in the smoky coal subpopulation.

Comparison of Exposure-Specific Models to Models Based on Coal Deposit Layers

We used ROC curve analysis to determine whether exposure-based models provided better discrimination between cases and controls than models based only on information on most frequently used coal deposit layers (see Figure S1). ROC curves and AUCs are given in Figure 4A, and they show that models based on cumulative exposure to individual pollutants provided better model predictions than models based on coal deposit information only (AUC of 79% vs. 72%). In contrast, predictive performance as measured by AUC improved only slightly when information on coal deposits was added to a model based on cumulative exposure to individual pollutants only. However, a comparison of model predictions for this combined model to that from the model based only on individual pollutant information indicated that the modeled pollutants could not fully predict lung cancer case–control status in subjects using coal from deposit 9 (Figure 4B), as evidenced by the symbols for this deposit visibly being off of the diagonal. This deposit was associated with the very highest risk of lung cancer in previous analyses (see also Figure S1) (Wong et al. 2019), and the estimated residual OR for subjects using coal from this deposit was estimated to be 2.77 (95% CI: 1.12, 6.85) in a model that also accounted for individual exposure effects. By contrast, the ORs for the other deposits ranged from 0.84 to 1.38, and all had 95% CIs crossing 1 (Table S12).

For the smoky coal subpopulation, ROC analysis (Figure 4C,D) showed that the exposure-based model (AUC = 72%) provided

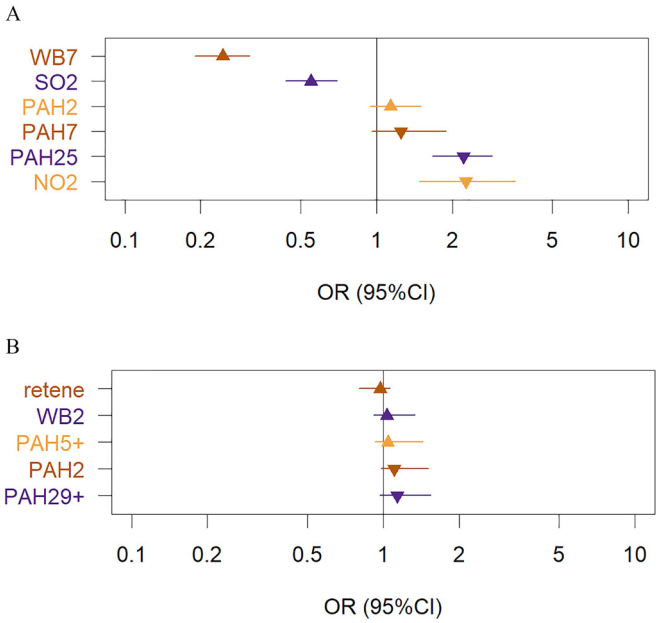

Figure 1. Odds ratios (ORs) per 1 standard deviation increase in exposure for different exposure clusters in the full population (A) and the smoky coal subpopulation (B). Exposure clusters were polycyclic aromatic hydrocarbon (PAH) 25 [consisting of 25 PAHs, including benzo(*a*)pyrene (BaP) and 5-methylchrysene (5-MC)], WB7 [consisting of seven wood burning–associated exposures, including fine particulate matter with aerodynamic diameter $\leq$2.5 μm ($PM_{2.5}$)], PAH7, PAH2, and two clusters consisting of the single exposures, nitrogen dioxide (NO_2) and sulfur dioxide (SO_2), respectively (See also Table S2 for complete listing of compounds in each cluster). Note: CI, confidence interval.

even larger improvements over a model based on coal deposit information only (AUC = 57%), while again indicating that the effect may be underestimated for subjects using coal from deposit 9 (OR: 1.58; 95% CI: 0.91, 4.11; Table S13). Overall, these analyses show that modeled exposures were able to explain much of the lung cancer likelihood attributable to specific types of coal in this region, but that the very highest-risk coal and its combustion products may have additional carcinogenic characteristics that remain to be identified.

Discussion

We conducted the largest population-based case–control study of lung cancer among never-smoking females in Xuanwei and Fuyuan, China, to date to identify the coal combustion emission constituents that drive the exceptionally high lung cancer rates in this region. A parallel comprehensive exposure assessment study, designed to sample coal deposit types and patterns of use representative of cases and controls, measured coal combustion exposures in this region. This data was then used to estimate lifelong exposure to an extensive number of compounds that are potentially toxic constituents of coal combustion in this region of China as well as standard constituents of air pollution. We identified that lung cancer was most strongly associated with two factors, namely, a cluster of 25 PAHs and NO_2.

Research to date on the cause of the extraordinarily high lung cancer rate in this region has identified the use of a particular type of bituminous coal, which was formed during the late Permian geologic time period, as the main driver of the observed risk (Large et al. 2009). However, studies to date have not been able to identify which HAP constituents are responsible for the observed risk. Understanding which individual or groups of constituents directly contribute to the development of lung cancer in this area is of great importance to further our understanding of lung cancer etiology.

In a previous analysis, we have shown that the association between the domestic use of smoky coal in Xuanwei and Fuyuan could vary from a twofold to a more than 30-fold increase in risk depending on the coal deposit sourced (Wong et al. 2019). In the current study, we show that this is largely driven by a cluster of parent PAHs and substituted derivatives, which include methylated PAHs. We attempted to isolate the signal of individual HAP components by employing a regularized regression model and identified 5-MC, benz(*a*)anthracene, and NO_2 as having the highest individual ORs in both the total study population and the smoky coal subpopulation.

Previously, Mumford et al. (1989, 1995) conducted a series of experimental studies of smoky coal combustion constituents from Xuanwei, reported that alkylated PAHs (which included 5-MC) are major mutagens in Xuanwei indoor air, and proposed that these may explain the very high rates of lung cancer in this region. There are several carcinogenic metabolites in 5-MC, and it is associated with malignancies in laboratory studies and is present in tobacco smoke and diesel soot extracts (Hecht et al. 1974, 1985; Hutzler et al. 2011; Pataki et al. 1983; Richter-Brockmann and Achten 2018). In addition, a recent study of coal from other parts of China showed that the method of preparation prior to combustion can have a drastic impact on the generation of methylated PAHs (Chen et al. 2015). The consistency in findings between the current paper and previous experimental studies suggests that methylated PAHs (including 5-MC) may play an important role in the very high lung cancer rates in this region. However, we should note that the high internal correlation within the PAH25 cluster means that identifying any single variable as the sole responsible variable is extremely challenging and that the high loading on 5-MC may represent a statistical, instead of a biological, effect.

We found that NO_2 was associated with lung cancer independently of other variables, both in the cluster and regularized regression models. The observed association with NO_2 is of interest, as studies of outdoor air pollution have provided consistent evidence of a relationship between NO_2, which is not thought to be a carcinogen itself, as a proxy for traffic-sourced air pollution exposure and lung cancer (Hamra et al. 2015). Similarly, it is possible that here, NO_2 is acting as a surrogate for a HAP component, which is relevant for lung cancer risk but is not reflected in any other constituents we have measured.

Previous research has proposed alternative potential drivers for the lung cancer risk in this region. Studies of uncombusted coal and controlled burning experiments have identified crystalline silica/quartz, either working alone or paired with volatile compounds, as a potential etiological agent. Silica is a known lung carcinogen, largely identified as such within the occupational setting (Steenland et al. 2001). However, it is unlikely that silica can explain a substantial part of the large lung cancer excess in Xuanwei and Fuyuan, given that lung cancer risk associated with high occupational silica exposure, where levels are two orders of magnitude higher than here, is only elevated by 50% (Vermeulen et al. 2011). Exposure to BaP has also been implicated in the region as an exposure of note (Mumford et al. 1987). BaP is well recognized as a lung carcinogen and has frequently been identified to be present in high amounts in the HAP of smoky coal homes, including in our exposure assessment study (Downward et al. 2014b). However, BaP was also observed to be in comparable amounts in wood-burning homes, indicating that its role as a driver of lung cancer in Xuanwei may be more limited. This is borne out by our analysis, which, in regularized regression, finds null ORs for BaP in both the full and smoky

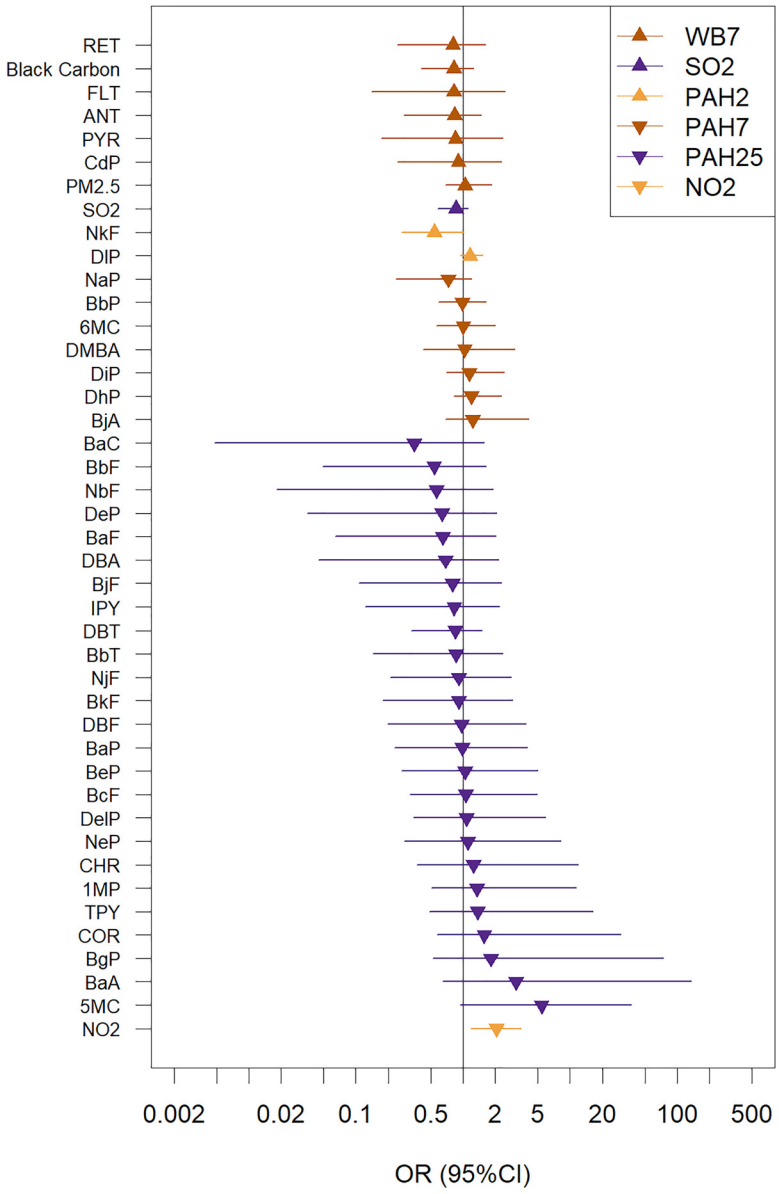

Figure 2. Odds ratios per 1 standard deviation increase in exposure for individual pollutants and lung cancer in the full population.

coal populations. Other components of coal combustion, which could potentially contribute to lung cancer, include elemental composition of HAP and the presence of radon in the air. We previously measured multiple elements in coal and HAP samples including arsenic, vanadium, and lead, and found that the elemental composition of HAP was not related to fuel or stove type (Downward et al. 2014a; Downward 2015). Radon measurements were collected from 57 households with an average radon concentration of 3.4 pCi/L and with no evidence of a difference between smoky and smokeless coal. The average detected radon level was below the action level of 4 pCi/L as set by the World Health Organization and corresponds with only an approximate 20% increase in lung cancer [National Research Council (U.S.) Committee on Evaluation of EPA Guidelines for Exposure to Naturally Occurring Radioactive Materials 1999], and, as with silica, this is not enough to explain the large excess risk of lung cancer in this region.

We employed ROC analysis to assess how well the different models discriminate between cases and controls in this study under different modeling circumstances. In general, the use of cumulative HAP constituent exposure information substantially improved model predictions vs. models using information only on coal deposits. However, our models were unable to predict the entirety of the cancer risk in subjects using coal from deposit 9, which is located in central Laibin, where coal deposits are associated with the highest lung cancer rates in Xuanwei (Figure S1). This suggests that while we have identified many of the major agents linked to lung cancer risk in much of this region, there may be additional components of exposure within the central Laibin area that have not yet been fully characterized. In addition to exposures, gene–environment interactions may also play a role. For example, perturbations in the *GTSM1* gene (which has a role in PAH metabolism) has been related to increased lung cancer risk in Xuanwei (Lan et al. 2000).

Our study had notable strengths. First, we had high participation rates. Second, we used empirical models derived from a comprehensive exposure survey of middle-aged and elderly never-smoking women in this region to predict over 40 HAP constituents for participants across their life course. The ability to assign individual exposure levels across the life course is especially unique to

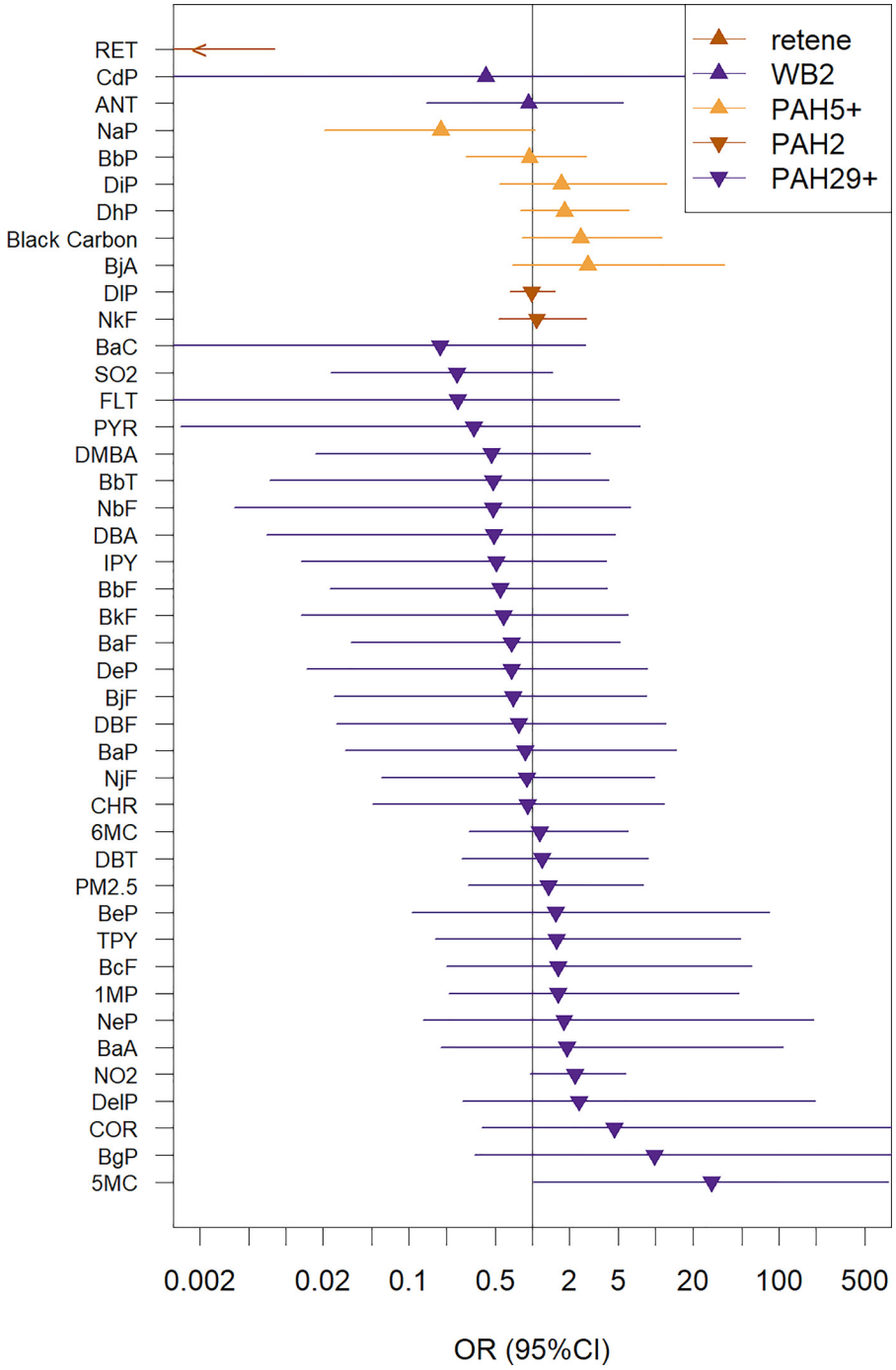

Figure 3. Odds ratios for individual pollutants and lung cancer in the smoky coal subpopulation. Estimates were rescaled to reflect the estimated effect for a -standard deviation increase in exposure in the full population to allow direct comparison with estimated effects in Figure 2. The (rescaled) point estimate for retene was off the scale (indicated by the "<" symbol).

this study, as no case–control study to date has been able to confidently link HAP concentrations to individual study participants. Many of the previous attempts to implicate groups or individual compounds to the observed risks were on an ecological level, which are prone to biases. Third, the study population was composed of Chinese women who never smoked, which removes potential confounding by active smoking and sex.

Our study has a number of limitations. Due to the high correlations between the different constituents, especially among PAHs and methylated PAHs, we cannot discount the possibility that other constituents also contribute to the observed ORs but, due to chance or having a relatively larger measurement error, were not retained in the models. Additionally, the high correlations between the PAHs, including those with strong negative effects (e.g., retene), are likely to result in effect estimates and their uncertainties being inflated, which is observed with the very wide uncertainties of the regularized regression among the smoky coal subpopulation.

An additional limitation is related to the life course approach. As with many case–control studies, recall bias is possible, especially when recalling events over the life course. Additionally, residential and household changes raise the potential for exposure misclassifications. Information regarding changes in residence,

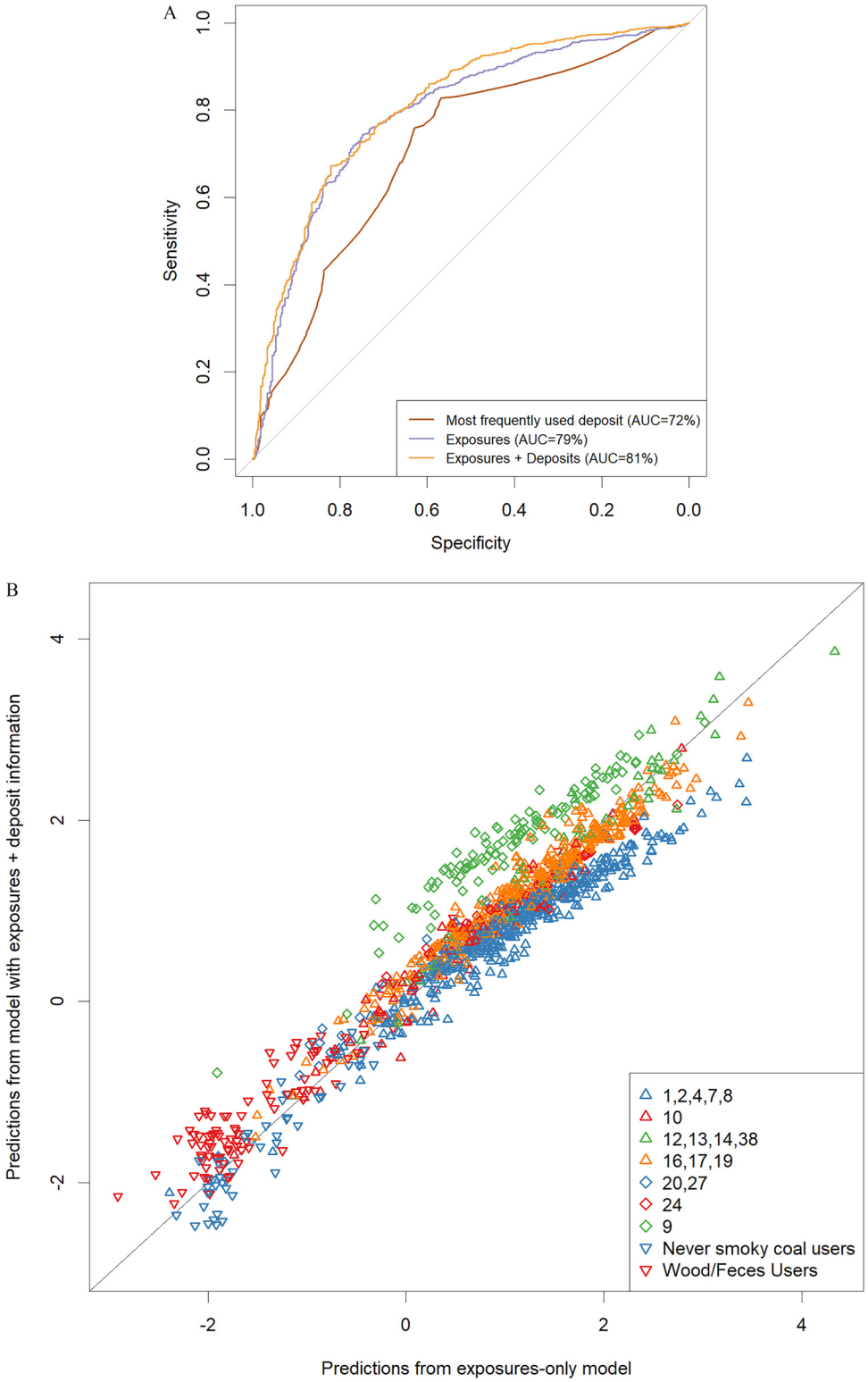

Figure 4. Receiver operating characteristic (ROC) curve analyses and model prediction quality of logistic models for household air pollutant (HAP) exposures vs. coal deposit layers and risk of lung cancer. Models were fitted to data from the full study population (A,B) or the smoky coal subpopulation (C,D). (A,C) show a comparison of the area under the curve (AUC) for lung cancer case and control predictions from a model including only the exposure estimates, a model including only information on the most frequently used deposit, and a model including both. ROC curves and corresponding AUC were estimated using cross validation. (B,D) show a comparison of model predictions for predictions based on an exposures-only model vs. a model with exposures + deposit information. The location of the referenced deposits is indicated in Figure S1. Predictions for deposit information is derived from logistic regression–adjusted forage and food sufficiency before marriage or 20 y of age and reported by Wong et al. (2019).

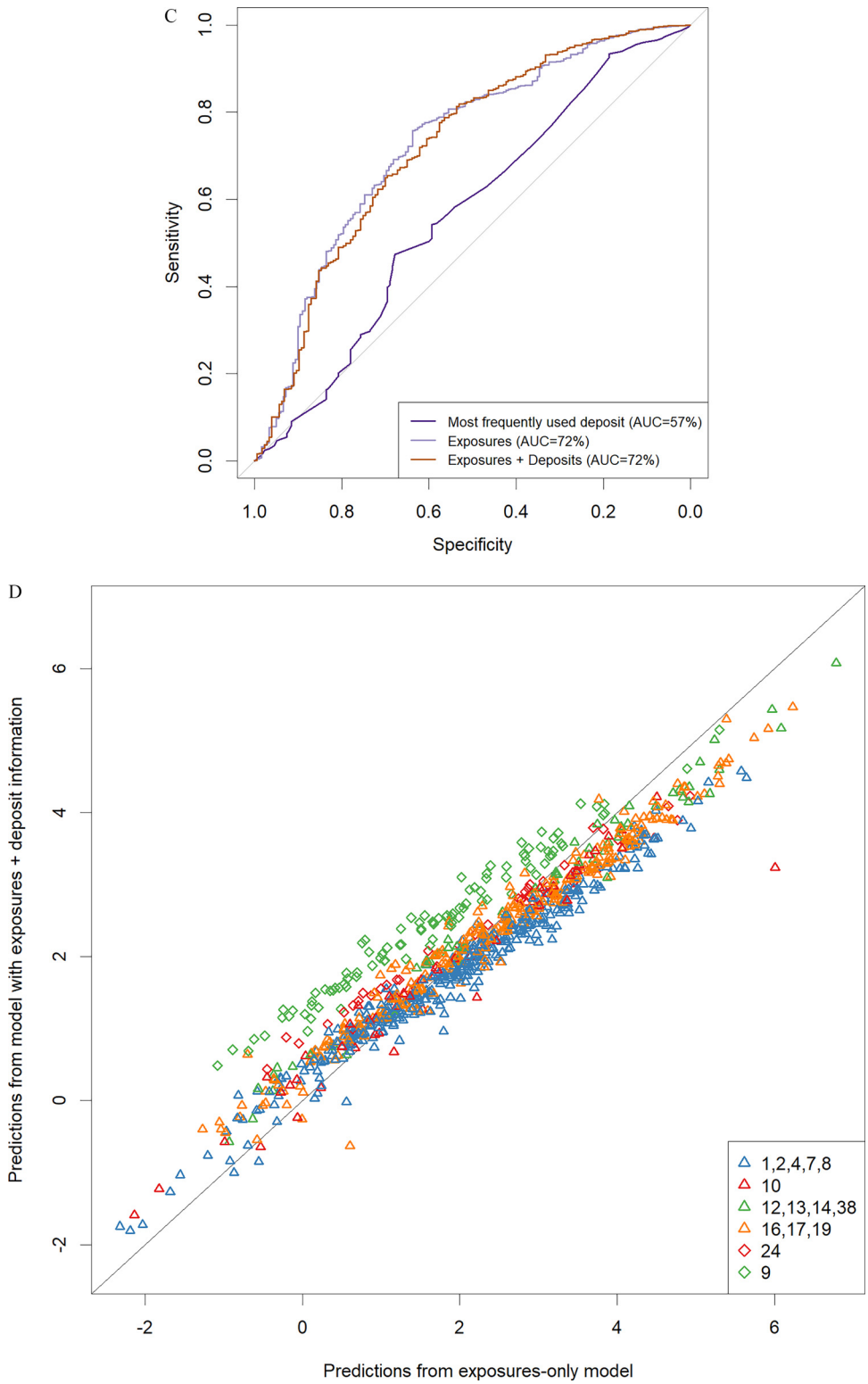

Figure 4. (*Continued.*)

coal sources, and exposure-relevant changes in household characteristics (e.g., chimneys) were collected as part of the current study and were included in the exposure models. Residents of Xuanwei and Fuyuan typically have limited residential mobility; therefore, events such as changes in address and the installation of improved stoves represent major (and therefore memorable to both cases and controls) life events. Additionally, measurements in the exposure study were collected in 2008–2009, while the exposures relevant to cancer risk likely occurred decades prior. To address this issue, we analyzed coal samples collected in the late 1980s from representative coal deposits in Xuanwei and coal samples collected from the same deposits in 2008–2013 and found that coal samples collected from these eras were comparable (Downward 2015; Large et al. 2009).

In summary, this is the first analytic epidemiologic study to comprehensively study the role of specific HAP constituents experienced over the entire life course and lung cancer risk. Cluster analysis indicated that the main drivers of lung cancer were NO_2 and a cluster of 25 PAHs. Within the PAH cluster, was some evidence to implicate 5-MC as a potential driver of lung cancer, which is especially relevant given that 5-MC is part of the family of methylated PAHs that have been found experimentally to be the major class of mutagens in combustion products of smoky coal in Xuanwei. Given the ubiquity of coal use for indoor cooking and heating in China and other countries and for electric power generation worldwide, our study suggests that more extensive monitoring of coal combustion products, including methylated PAHs, may be warranted to more accurately assess health risks from this exposure. Finally, a fuller understanding of the presence and determinants of production of these compounds is needed to minimize exposure, which could have important public health and prevention implications in China and elsewhere.

Acknowledgments

This study was supported with intramural funding from the National Cancer Institute. We thank J. King, P. Hui, N. Appel, and L. Carroll for their support. We dedicate this work to the memory of R. Chapman and X. He, outstanding mentors and pioneers in the study of the lung cancer excess in this part of the world.

References

Agier L, Portengen L, Chadeau-Hyam M, Basagaña X, Giorgis-Allemand L, Siroux V, et al. 2016. A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. Environ Health Perspect 124(12):1848–1856, PMID: 27219331, https://doi.org/10.1289/EHP172.

Barone-Adesi F, Chapman RS, Silverman DT, He X, Hu W, Vermeulen R, et al. 2012. Risk of lung cancer associated with domestic use of coal in Xuanwei, China: retrospective cohort study. BMJ 345:e5414, PMID: 22936785, https://doi.org/10.1136/bmj.e5414.

Betancourt M. 2018. *A Conceptual Introduction to Hamiltonian Monte Carlo*. https://arxiv.org/pdf/1701.02434.pdf, [accessed 15 December 2018].

Box GEP, Cox DR. 1964. An analysis of transformations. J Roy Stat Soc B 26:211–243, https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Brooks SP, Gelman A. 1998. General methods for monitoring convergence of iterative simulations. J Comput Graph Stat 7(4):434–455, https://doi.org/10.2307/1390675.

Bürkner P. 2017. Brms: an R package for Bayesian multilevel models using stan. J Stat Soft 80(1):1–27, https://doi.org/10.18637/jss.v080.i01.

Carvalho CM, Polson NG, Scott JG. 2010. The horseshoe estimator for sparse signals. Biometrika 97(2):465–480, https://doi.org/10.1093/biomet/asq017.

Casella G, Ghosh M, Gill J, Kyung M. 2010. Penalized regression, standard errors, and Bayesian lassos. Bayesian Anal 5(2):369–411, https://doi.org/10.1214/10-BA607.

CDC (Centers for Disease Control and Prevention). 2013. *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM). http://www.cdc.gov/nchs/icd/icd9cm.htm, [accessed 29 August 2019].

Chapman RS, Mumford JL, Harris DB, He ZZ, Jiang WZ, Yang RD. 1988. The epidemiology of lung cancer in Xuan Wei, China: current progress, issues, and research strategies. Arch Environ Health 43(2):180–185, PMID: 3377554, https://doi.org/10.1080/00039896.1988.9935850.

Chen P. 2000. Study on integrated classification system for Chinese coal. Fuel Process Technol 62(2–3):77–87, https://doi.org/10.1016/S0378-3820(99)00115-0.

Chen Y, Zhi G, Feng Y, Chongguo T, Bi X, Li J, et al. 2015. Increase in polycyclic aromatic hydrocarbon (PAH) emissions due to briquetting: a challenge to the coal briquetting policy. Environ Pollut 204:58–63, PMID: 25912887, https://doi.org/10.1016/j.envpol.2015.04.012.

Couraud S, Zalcman G, Milleron B, Morin F, Souquet PJ. 2012. Lung cancer in never smokers–a review. Eur J Cancer 48(9):1299–1311, PMID: 22464348, https://doi.org/10.1016/j.ejca.2012.03.007.

Defays D. 1977. Efficient algorithm for a complete link method. Comput J 20(4):364–366, https://doi.org/10.1093/comjnl/20.4.364.

Downward GS. 2015. Quantification and characterization of household air pollution exposure from the use of solid fuels; clues to the lung cancer epidemic in Xuanwei and Fuyuan. Ph.D. thesis, China: Utrecht University. https://dspace.library.uu.nl/handle/1874/312312, [accessed 1 May 2019].

Downward GS, Hu W, Large D, Veld H, Xu J, Reiss B, et al. 2014a. Heterogeneity in coal composition and implications for lung cancer risk in Xuanwei and Fuyuan counties, China. Environ Int 68:94–104, PMID: 24721117, https://doi.org/10.1016/j.envint.2014.03.019.

Downward GS, Hu W, Rothman N, Reiss B, Tromp P, Wu G, et al. 2017. Quartz in ash, and air in a high lung cancer incidence area in China. Environ Pollut 221:318–325, PMID: 27939206, https://doi.org/10.1016/j.envpol.2016.11.081.

Downward GS, Hu W, Rothman N, Reiss B, Wu G, Wei F, et al. 2014b. Polycyclic aromatic hydrocarbon exposure in household air pollution from solid fuel combustion among the female population of Xuanwei and Fuyuan counties, China. Environ Sci Technol 48(24):14632–14641, PMID: 25393345, https://doi.org/10.1021/es504102z.

Downward GS, Hu W, Rothman N, Reiss B, Wu G, Wei F, et al. 2016. Outdoor, indoor, and personal black carbon exposure from cookstoves burning solid fuels. Indoor Air 26(5):784–795, PMID: 26452237, https://doi.org/10.1111/ina.12255.

Epplein M, Schwartz SM, Potter JD, Weiss NS. 2005. Smoking-adjusted lung cancer incidence among Asian-Americans (United States). Cancer Causes Control 16(9):1085–1090, PMID: 16184474, https://doi.org/10.1007/s10552-005-0330-6.

Fawcett T. 2006. An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874, https://doi.org/10.1016/j.patrec.2005.10.010.

Gelman A, Jakulin A, Pittau MG, Su YS. 2008. A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat 2(4):1360–1383, https://doi.org/10.1214/08-AOAS191.

Hamra GB, Laden F, Cohen AJ, Raaschou-Nielsen O, Brauer M, Loomis D. 2015. Lung cancer and exposure to nitrogen dioxide and traffic: a systematic review and meta-analysis. Environ Health Perspect 123(11):1107–1112, PMID: 25870974, https://doi.org/10.1289/ehp.1408882.

Hecht SS, Bondinell WE, Hoffmann D. 1974. Chrysene and methylchrysenes: presence in tobacco smoke and carcinogenicity. J Natl Cancer Inst 53(4):1121–1133, PMID: 4427391, https://doi.org/10.1093/jnci/53.4.1121.

Hecht SS, Radok L, Amin S, Huie K, Melikian AA, Hoffmann D, et al. 1985. Tumorigenicity of 5-methylchrysene dihydrodiols and dihydrodiol epoxides in newborn mice and on mouse skin. Cancer Res 45(4):1449–1452, PMID: 3838497.

Ho KF, Chang CC, Tian L, Chan CS, Musa Bandowe BA, Lui KH, et al. 2016. Effects of polycyclic aromatic compounds in fine particulate matter generated from household coal combustion on response to EGFR mutations in vitro. Environ Pollut 218:1262–1269, PMID: 27613327, https://doi.org/10.1016/j.envpol.2016.08.084.

Hu W, Downward GS, Reiss B, Xu J, Bassig BA, Hosgood HD, 3rd, et al. 2014. Personal and indoor PM2.5 exposure from burning solid fuels in vented and unvented stoves in a rural region of China with a high incidence of lung cancer. Environ Sci Technol 48(15):8456–8464, PMID: 25003800, https://doi.org/10.1021/es502201s.

Hutzler C, Luch A, Filser JG. 2011. Analysis of carcinogenic polycyclic aromatic hydrocarbons in complex environmental mixtures by LC-APPI-MS/MS. Anal Chim Acta 702(2):218–224, PMID: 21839201, https://doi.org/10.1016/j.aca.2011.07.003.

Lan Q, He X, Costa DJ, Tian L, Rothman N, Hu G, et al. 2000. Indoor coal combustion emissions, GSTM1 and GSTT1 genotypes, and lung cancer risk: a case-control study in Xuan Wei, China. Cancer Epidemiol Biomarkers Prev 9(6):605–608, PMID: 10868696.

Lan Q, He X, Shen M, Tian L, Liu LZ, Lai H, et al. 2008. Variation in lung cancer risk by smoky coal subtype in Xuanwei, China. Int J Cancer 123(9):2164–2169, PMID: 18712724, https://doi.org/10.1002/ijc.23748.

Large DJ, Kelly S, Spiro B, Tian L, Shao L, Finkelman R, et al. 2009. Silica-volatile interaction and the geological cause of the Xuan Wei lung cancer epidemic. Environ Sci Technol 43(23):9016–9021, PMID: 19943682, https://doi.org/10.1021/es902033j.

Lui KH, Bandowe BA, Tian L, Chan CS, Cao JJ, Ning Z, et al. 2017a. Cancer risk from polycyclic aromatic compounds in fine particulate matter generated from household coal combustion in Xuanwei, China. Chemosphere 169:660–668, PMID: 27912191, https://doi.org/10.1016/j.chemosphere.2016.11.112.

Lui KH, Dai WT, Chan CS, Tian L, Ning BF, Zhou Y, et al. 2017b. Cancer risk from gaseous carbonyl compounds in indoor environment generated from household coal combustion in Xuanwei, China. Environ Sci Pollut Res Int 24(21):17500–17510, PMID: 28593548, https://doi.org/10.1007/s11356-017-9223-y.

Mumford JL, Chapman RS, Harris DB, He XZ, Cao SR, Xian YL, et al. 1989. Indoor air exposure to coal and wood combustion emissions associated with a high lung-cancer rate in Xuan Wei, China. Environ Int 15(1–6):315–320, https://doi.org/10.1016/0160-4120(89)90044-5.

Mumford JL, He XZ, Chapman RS, Cao SR, Harris DB, Li XM, et al. 1987. Lung cancer and indoor air pollution in Xuan Wei, China. Science 235(4785):217–220, PMID: 3798109, https://doi.org/10.1126/science.3798109.

Mumford JL, Li X, Hu F, Lu XB, Chuang JC. 1995. Human exposure and dosimetry of polycyclic aromatic hydrocarbons in urine from Xuan Wei, China with high

lung cancer mortality associated with exposure to unvented coal smoke. Carcinogenesis 16(12):3031–3036, PMID: 8603481, https://doi.org/10.1093/carcin/16.12.3031.

National Bureau of Statistics of China. 2000. Population of townships, streets in Yunnan province. http://www.stats.gov.cn/tjsj/ndsj/renkoupucha/2000jiedao/html/J53.htm [accessed 7 December 2019].

National Research Council (U.S.) Committee on Evaluation of EPA Guidelines for Exposure to Naturally Occurring Radioactive Materials. 1999. Indoor-radon guidelines and recommendations. In: *Evaluation of Guidelines for Exposures to Technologically Enhanced Naturally Occurring Radioactive Materials*. Washington, DC: National Academies Press.

Navidi W, Thomas D, Stram D, Peters J. 1994. Design and analysis of multilevel analytic studies with applications to a study of air pollution. Environ Health Perspect 102(Suppl 8):25–32, PMID: 7851327, https://doi.org/10.1289/ehp.94102s825.

Pataki J, Lee H, Harvey RG. 1983. Carcinogenic metabolites of 5-methylchrysene. Carcinogenesis 4(4):399–402, PMID: 6839413, https://doi.org/10.1093/carcin/4.4.399.

Pelosof L, Ahn C, Gao A, Horn L, Madrigales A, Cox J, et al. 2017. Proportion of never-smoker non-small cell lung cancer patients at three diverse institutions. J Natl Cancer Inst 109(7):djw295, PMID: 28132018, https://doi.org/10.1093/jnci/djw295.

Piironen J, Vehtari A. 2017. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). 20–22 April 2017*. Fort Lauderdale, FL: AISTATS, 905–913.

Richter-Brockmann S, Achten C. 2018. Analysis and toxicity of 59 PAH in petrogenic and pyrogenic environmental samples including dibenzopyrenes, 7H-benzo[c]fluorene, 5-methylchrysene and 1-methylpyrene. Chemosphere Jun 200:495–503, PMID: 29505926, https://doi.org/10.1016/j.chemosphere.2018.02.146.

Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65, https://doi.org/10.1016/0377-0427(87)90125-7.

Samet JM, Avila-Tang E, Boffetta P, Hannan LM, Olivo-Marston S, Thun MJ, et al. 2009. Lung cancer in never smokers: clinical epidemiology and environmental risk factors. Clin Cancer Res 15(18):5626–5645, PMID: 19755391, https://doi.org/10.1158/1078-0432.CCR-09-0376.

Schisterman EF, Perkins NJ, Mumford SL, Ahrens KA, Mitchell EM. 2017. Collinearity and causal diagrams: a lesson on the importance of model specification. Epidemiology 28(1):47–53, PMID: 27676260, https://doi.org/10.1097/EDE.0000000000000554.

Seow WJ, Downward GS, Wei H, Rothman N, Reiss B, Xu J, et al. 2016. Indoor concentrations of nitrogen dioxide and sulfur dioxide from burning solid fuels for cooking and heating in Yunnan Province, China. Indoor Air 26(5):776–783, PMID: 26340585, https://doi.org/10.1111/ina.12251.

Sisti J, Boffetta P. 2012. What proportion of lung cancer in never-smokers can be attributed to known risk factors? Int J Cancer 131(2):265–275, PMID: 22322343, https://doi.org/10.1002/ijc.27477.

Steenland K, Mannetje A, Boffetta P, Stayner L, Attfield M, Chen J, et al. 2001. Pooled exposure-response analyses and risk assessment for lung cancer in 10 cohorts of silica-exposed workers: an IARC multicentre study. Cancer Causes Control 12(9):773–784, PMID: 11714104.

Tian L, Lucas D, Fischer SL, Lee SC, Hammond SK, Koshland CP. 2008. Particle and gas emissions from a simulated coal-burning household fire pit. Environ Sci Technol 42(7):2503–2508, PMID: 18504988, https://doi.org/10.1021/es0716610.

Tian L. 2005. Coal combustion emissions and lung cancer in Xuan Wei, China. Ph.D. thesis, CA: University of California, Berkeley. https://search.proquest.com/openview/95bafa5279327af1fe434b968204de2b/1?pq-origsite=gscholar&cbl=18750&diss=y.

Vermeulen R, Rothman N, Lan Q. 2011. Coal combustion and lung cancer risk in Xuanwei: a possible role of silica? Med Lav 102(4):362–367, PMID: 21834273.

Weisskopf MG, Seals RM, Webster TF. 2018. Bias amplification in epidemiologic analysis of exposure to mixtures. Environ Health Perspect 126(4):047003, PMID: 29624292, https://doi.org/10.1289/EHP2450.

Wong JYY, Downward GS, Hu W, Portengen L, Seow WJ, Silverman DT, et al. 2019. Lung cancer risk by geologic coal deposits: a case-control study of female never-smokers from Xuanwei and Fuyuan, China. Int J Cancer 144(12):2918–2927, PMID: 30511435, https://doi.org/10.1002/ijc.32034.

Yu T, Li J, Ma S. 2012. Adjusting confounders in ranking biomarkers: a model-based ROC approach. Brief Bioinform 13(5):513–523, PMID: 22396461, https://doi.org/10.1093/bib/bbs008.