Environment International 173 (2023) 107870

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Full length article

Methylated polycyclic aromatic hydrocarbons from household coal use across the life course and risk of lung cancer in a large cohort of 42,420 subjects in Xuanwei, China

Lützen Portengen [a,*], George Downward [a], Bryan A. Bassig [a,1], Batel Blechter [b], Wei Hu [b], Jason Y.Y. Wong [b], Bofu Ning [c], Mohammad L. Rahman [b], Bu-Tian Ji [b], Jihua Li [d], Kaiyun Yang [e], H. Dean Hosgood [f], Debra T. Silverman [b], Nathaniel Rothman [b], Yunchao Huang [g,2], Roel Vermeulen [a,2], Qing Lan [b,2]

[a] Institute for Risk Assessment Sciences, Utrecht University, Utrecht, Utrecht, the Netherlands
[b] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA
[c] Xuanwei Center for Disease Control and Prevention, Xuanwei, Qujing, Yunnan, China
[d] Qujing Center for Diseases Control and Prevention, Sanjiangdadao, Qujing, Yunnan, China
[e] Third Affiliated Hospital of Kunming Medical University (Yunnan Tumor Hospital), Kunming, China
[f] Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, NY, USA
[g] Department of Epidemiology, Shanghai Cancer Institute, Shanghai Jiaotong University, Shanghai, China

ARTICLE INFO

Keywords:
Lifecourse epidemiology
Indoor air pollution
Lung cancer epidemiology
Solid biomass fuel burning
Polycyclic aromatic hydrocarbons

ABSTRACT

Background: We previously showed that exposure to 5-methylchrysene (5MC) and other methylated polycyclic aromatic hydrocarbons (PAHs) best explains lung cancer risks in a case-control study among non-smoking women using smoky coal in China. Time-related factors (e.g., age at exposure) and non-linear relations were not explored.
Objective: We investigated the relation between coal-derived air pollutants and lung cancer mortality using data from a large retrospective cohort.
Methods: Participants were smoky (bituminous) or smokeless (anthracite) coal users from a cohort of 42,420 subjects from four communes in XuanWei. Follow-up was from 1976 to 2011, during which 4,827 deaths from lung-cancer occurred. Exposures were predicted for 43 different pollutants. Exposure clusters were identified using hierarchical clustering. Cox regression was used to estimate exposure–response relations for 5MC, while effect modification by age at exposure was investigated for cluster prototypes. A Bayesian penalized multi-pollutant model was fitted on a nested case-control sample, with more restricted models fitted to investigate non-linear exposure–response relations.
Results: We confirmed the strong exposure–response relation for 5MC (Hazard Ratio [95% Confidence Interval] = 2.5 [2.4, 2.6] per standard-deviation (SD)). We identified four pollutant clusters, with all but two PAHs in a single cluster. Exposure to PAHs in the large cluster was associated with a higher lung cancer mortality rate (HR [95%CI] = 2.4 [2.2, 2.6] per SD), while exposure accrued before 18 years of age appeared more important than adulthood exposures. Results from the multi-pollutant model identified anthanthrene (ANT) and benzo(a) chrysene (BaC) as risk factors. 5MC remained strongly associated with lung cancer in models that included ANT and BaC and also benzo(a)pyrene (BaP).
Conclusion: We confirmed the link between PAH exposures and lung cancer in smoky coal users and found exposures before age 18 to be especially important. We found some evidence for the carcinogen 5MC and non-carcinogens ANT and BaC.

* Corresponding author at: Postbus 80178, 3508 TD Utrecht, the Netherlands.
E-mail address: l.portengen@uu.nl (L. Portengen).
[1] This author is now employed by the U.S. Centers for Disease Control and Prevention, National Center for Health Statistics. All work was conducted while employed by the National Cancer Institute.
[2] These authors co-supervised this work.

https://doi.org/10.1016/j.envint.2023.107870
Received 27 September 2022; Received in revised form 3 March 2023; Accepted 6 March 2023
Available online 7 March 2023

1. Introduction

Several studies have confirmed the relation between household air pollution (HAP), due to burning coals and other solid fuels, and mortality from respiratory diseases, including lung cancer. (Balmes, 2019; Gordon et al., 2014; Guan et al., 2016; IARC, 2016) Use of bituminous ("smoky") coal has been linked to the excess risk of lung cancer that was observed earlier in non-smoking women living in rural Xuanwei, China. (Lan et al., 2008; Barone-Adesi et al., 2012; Wong et al., 2018) Residents of this area primarily work as subsistence farmers and burn solid fuels for cooking and heating. Using data from a population-based case-control study of lung cancer among never-smoking women in Xuanwei and neighbouring Fuyuan, we showed previously that lung cancer incidence was strongly associated with cumulative exposure to methylated polycyclic aromatic hydrocarbons (PAHs) and NO_2, and that the evidence was strongest for exposure to 5-methylchrysene (5MC) (Vermeulen et al., 2019) rather than for the "known" carcinogens benzo[a]pyrene (BaP) and particulate matter (PM). (IARC, 2016)

The importance of using a life course approach, studying the temporal relation between air pollution exposure during gestation, childhood, adolescence, or (young) adulthood and subsequent chronic disease, has been stressed in a number of publications. (Lee et al., 2015; Phillips et al., 2018) There is evidence that exposures at younger ages may have more detrimental effects due to the still developing organ systems. (Belgrave et al., 2018; Bui et al., 2018; Liu and Kalhan, 2021) Case-control studies can be used to study life course questions, (Maheswaran et al., 2002) but control selection can be a challenge and birth cohort studies are therefore often preferred.

Our earlier investigation of exposure–response relations for constituents of household air pollution and lung cancer, relied on models that focused on the linear effects of cumulative exposure on the log-odds of lung cancer, although we reported results also from an analysis where exposures had been categorized. This assumption of a linear response can be restrictive when the true exposure response is not linear, and may result in incorrect ranking of mixture components. (Lazarevic et al., 2020) In addition, non-linear effects may be of interest from an etiological or risk assessment perspective (Vermeulen et al., 2014; Vlaanderen et al., 2010).

"To follow-up on our previous case-control study that found a strong relation between exposure to coal-derived methylated PAHs and lung cancer incidence, we investigated the relation between the same set of household air pollutants and lung cancer mortality using data from a large cohort study of men and women living in Xuanwei, China. We additionally evaluated the importance of time-related factors (age at exposure) and potential non-linear exposure–response relations.

2. Methods

2.1. Population

2.1.1. Full cohort and lifetime smoky-coal user sub-cohort

The design and study population of this cohort have been previously described. (Barone-Adesi et al., 2012; Lan et al., 2002) Briefly, subjects were sampled from a rural population in Xuanwei, China, who were still primarily using coal for indoor heating and cooking. The study area comprised of three communes where residents mainly used a bituminous ("smoky") coal from a single coal seam and one commune where residents mainly used an anthracite ("smokeless") coal. A summary of subject characteristics by commune is provided in the appendix (table S1). The retrospective part of the cohort was assembled in 1992 and was based on a review of administrative records to identify all subjects born between 1917 and 1951 who lived in the study area as of 1 January 1976. A graphical illustration of the timeline of the study is provided in the appendix (figure S1). The full cohort included 42,420 subjects, but after excluding subjects with missing or unreliable information on vital status (n = 277) or coal use (n = 60), 42,083 subjects remained for further analyses. Of these, 29,032 reported using coal only from deposits that were regarded as smoky coal deposits (i.e., lifetime smoky coal users). Note that this definition of a lifetime smoky coal user is slightly different from that used in other papers on this population, (Wong et al., 2018) where the definition was based on self-reported coal type instead of self-reported coal source. However, self-reported coal type by participants from our exposure survey study did not always match the physico-chemical properties of the actual coal used (assessed using chemical analyses), and we therefore based our definition on the physico-chemical properties of coals from the seam where subjects reportedly sourced their coal. The remaining 13,051 subjects mainly used smokeless coal or other solid fuels (wood, plants etc.).

The study was approved by the Institutional Review Board of the Chinese Academy of Preventive Medicine. Informed consent was obtained from all participants or their proxies if they could not understand the written consent form.

2.1.2. Case ascertainment

The date and cause of death for participants in the cohort over the follow-up period (1 January 1976 to 31 December 2011) were obtained from hospital records and death certificates. A death certificate is required for official documentation in hospitals, public security bureaus, and public health bureaus in Xuanwei. Cause of death was coded by the Center for Disease Control according to the International Classification of Diseases, 9th revision (ICD-9), including for lung cancer (162).

A total of 4640 lung cancer deaths were identified.

2.1.3. Exposure assessment

Exposure assessment was performed similarly to the procedure used for our earlier case-control study (Vermeulen et al., 2019) and was described in detail for this cohort by Bassig et al. (Bassig et al., 2020) In short, self-reported information on fuel and stove use, including information on the mine from which coal was sourced, was available for each year from birth to end of follow-up or death from lung cancer. If subjects (or surrogate respondents) reported that they used coal from a local mine, the deposit layer was assigned the most likely deposit based on results from a Bayesian predictive model. The type of fuel and type of stove(s) that were used were the main determinants in determinant-based exposure models that were derived using household air pollutant measurements and determinant data collected during an exposure survey that was conducted in this population. (Downward et al., 2016; Downward et al., 2014; Hu et al., 2014) These exposure models were used to predict exposure levels for 39 different PAHs (a full list with the abbreviations used in this paper is provided in the appendix), as well as the known air pollutants NO_2, SO_2, particulate matter (PM), and black carbon (BC). Validity of model estimates with regards to long-term annual exposure was not evaluated due to the lack of historical measurements, but marginal R^2s for the exposure models were generally fair (in the range of 0.23–0.74). More detailed information regarding the exposure assessment and a table showing marginal R^2s for each exposure (table S2) are provided in the appendix.

2.1.4. Nested case-control sample

To facilitate high-dimensional cluster analyses of household air pollutant exposure data and to allow fitting of Bayesian penalized (horseshoe) regression models (that rely on computationally demanding Markov Chain Monte Carlo (MCMC) algorithms), a nested case-control sample was drawn from the full cohort. For each lung cancer case we drew 10 control subjects from participants that were alive at the age of death of the lung cancer case, matching on sex and year of birth. Subjects could be selected as controls for more than one case, and cases were considered to be "at risk" when still alive, meaning that they could be selected as controls for other cases.

2.2. Statistical analysis

2.2.1. Cluster analysis & derivation of cluster prototypes

In our main survival analyses, we used the so-called "counting process" formulation (Andersen and Gill, 1982) to account for yearly changes in exposure, which means that the analytical dataset contained multiple records for each individual (one for each age). This makes it more difficult to avoid double counting, while still summarizing the available exposure information in a comprehensive way. We therefore used data from the control subjects in our nested case-control samples to inform a hierarchical cluster analysis aimed at identifying clusters of exposures in the full cohort and the subset of lifetime smoky coal users respectively.

Exposures were transformed using power transformations to reduce skewness and improve symmetry of the distributions (Box and Cox, 1964). We used standard Euclidean distances and the complete linkage method to determine the cluster sequence and selected the number of clusters to extract based on the silhouette score (Rousseeuw, 1987). Based on the results from this analysis, we derived a cluster prototype scoring rule by extracting the first component score coefficients from a principal component analysis (PCA) model applied to the cluster-specific exposure data. Prototype scores were mean-centered and scaled based on data from the control subjects. Cluster scores in the full and lifetime smoky coal cohorts were calculated using the scoring and scaling rules derived from the nested case-control sample.

2.2.2. Survival analysis to assess cluster prototype effects using data from the full cohort

We used Cox regression from the *survival* package (Therneau, 2022) in R with age as the time scale, accounting for left truncation and yearly changes in (time-varying) exposure by using multiple records for each individual. Baseline rates were stratified by sex and year of birth and the models accounted for confounding by differences in smoking habits and educational status (as a proxy for socio-economic status).

2.2.3. Importance of timing of exposure

We previously found that women who began cooking with smoky coal $<$ 18 years of age had a higher risk of lung cancer compared to women who started cooking at an older age. (Barone-Adesi et al., 2012) To assess the impact of the timing of pollutant exposure we calculated cumulative exposure for the main PAH cluster prototype separately for the first 18 years and all later ages. We then evaluated, for each subject in each of the risk-sets used in the Cox analysis, whether they had been exposed either above or below (or equal to) the average cumulative exposure in that risk-set. Each risk-set is made up of a single lung cancer case and all sex- and birthyear-matched subjects in the cohort that were alive at the age that the case died as controls. Cumulative exposure was calculated separately before or after age 18 years until each risk-set's age. This metric was then used to estimate hazard ratios (HR) for being highly exposed on average before or after age 18 years. This metric was chosen to avoid comparing effect estimates for cumulative exposures accrued over very different exposure durations (i.e. 18 years for childhood exposure, but on average 37 years, range 9–72 years, for adult exposures) and to avoid interpretational problems that arise when subtracting values from a power-transformed variable.

2.2.4. Multi-pollutant models

Multi-pollutant modelling for the full set of 43 household air-pollutants was performed by fitting a (horseshoe) penalized conditional logistic regression model as implemented in the *brms* package (Bürkner, 2017) in R to data from the nested case-control sample(s). As in our previous paper, we followed recommendations for further modifications of the horseshoe prior and hyperparameter settings that were suggested by Piironen and Vehtari (Piironen and Vehtari, 2017). (Piironen and Vehtari, 2017) The hyperparameter for the scale of the global shrinkage parameter was selected assuming that all the confounders, but no more than 10% 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 (Carpenter et al., 2017) to 0.99 to avoid divergent transitions.

Household air-pollutant exposure levels were scaled before fitting the model, as is usual for penalized regression models, using scaling factors (empirical standard deviations) that were derived from the empirical exposure distributions in controls from each of the nested case-control samples (i.e., for the full cohort and for the lifetime smoky coal users). Scaling factors were used to re-scale effect estimates from models that were fitted on data with different scaling factors to make them comparable.

2.2.5. Non-linear exposure response relations

We fitted multi-pollutant Cox regression models that included a more restricted set of air pollutants, to data from the full cohort to assess the potential for non-linear exposure–response relations. These models used penalized splines as implemented in the *psline* function from the *survival* package (Therneau, 2022) in R and included only anthanthrene (ANT), benzo(a)chrysene (BaC), benzo(a)pyrene (BaP), and/or 5–methyl chrysene (5MC). The degrees of freedom for each spline function was selected based on Akaike's Information Criterion (AIC). (Akaike, 1974).

3. Results

A summary of subject characteristics is provided in Table 1. Almost all the male subjects (92%), but very few female subjects (<2%), smoked, making it impossible to investigate interaction between air pollutant exposure and smoking or effect modification by sex. Lifetime smoky coal users tended to be younger at entry into the cohort, were more often female and tended to be more educated than smokeless coal users or subjects that burnt mostly other fuels. There were fewer cases of lung cancer among the smokeless coal and other fuel users than among the lifetime smoky coal users and they tended to occur at a later age. A graph showing the much higher estimated age-specific lung cancer mortality rates for lifetime smoky coal users than for smokeless coal or other solid fuel users is provided in the appendix (figure S2).

A heatmap showing the correlation between individual pollutant exposures in controls from the nested case-control sample, both in the full cohort and the lifetime smoky coal subpopulation, is provided in the appendix (figure S3).

Table 1
Subject characteristics for lifetime smoky coal users and subjects using mostly smokeless coal or other solid fuels in a large population-based cohort from Xuanwei, China.

	Lifetime smoky coal users[a]	Mostly smokeless coal or other solid fuel users
Number of subjects	29,032	13,051
Person years of follow-up (years)	766,887	377,622
Age at entry, mean ± SD	38.3 ± 10.5	39.4 ± 10.8
Age at end of follow-up, mean ± SD	65.0 ± 10.7	68.5 ± 10.3
Number of lung cancer deaths	4,531	109
Age at lung cancer death, mean ± SD	59.7 ± 10.1	64.0 ± 9.6
Females, n (%)	14,339 (49.4)	6,209 (47.6)
Ever smoking, n (%)	14,162 (48.8)	6,650 (51.0)
Education, n (%)		
no formal education	18,483 (63.6)	8,881 (68.0)
primary school	8,846 (30.5)	3,521 (27.0)
middle school or higher	1,703 (5.9)	649 (5.0)

[a] Lifetime use status in this paper was based on self-reported coal source.

3.1. Replication of previous findings

A bi-pollutant model that included the two strongest risk factors (NO_2 and 5MC) identified in previous analyses of data from a (retrospective) case-control study in the same general area (Vermeulen et al., 2019) confirmed the much higher lung cancer mortality rates associated with exposure to 5MC (HR [95%CI] = 2.5 [2.4, 2.6]), but not for NO_2 (HR [95%CI] = 0.8 [0.7, 0.8]). The increase in lung cancer rates associated with 5MC exposure was much reduced when assessed in lifetime smoky coal users only (HR [95%CI] = 1.3 [1.2, 1.3]).

3.2. Exposure distributions and cluster analysis

We used household air pollutant data from the nested case-control sample to compare exposure distributions between cases and controls, with histograms provided in the appendix (figure S4). Exposure distributions for most of the PAHs tended to be bimodal in both cases and controls, but with relatively more low exposed controls and higher exposed cases. Exposure distributions for PM, BC, and NO_2 appeared more unimodal and not very different for cases and controls, while estimated exposures to SO_2 tended to be higher in controls than in cases.

In hierarchical cluster analysis using data from controls in the nested case-control sample of the full cohort, all but two of the 39 PAHs were assigned to a single cluster, reflecting the strong correlations between estimated PAH exposure levels in this population (appendix figure S3). PM, BC, and NO_2 also clustered together, while three pollutants (SO_2, NkF, and DlP) were each assigned their own (single-exposure) cluster. The optimal cluster solution for lifetime smoky coal subjects was very similar, with all but three PAHs assigned to a single large cluster, a second cluster that contained SO_2, PM, BC, and NO_2, and two clusters consisting of DlP and NkF and ANT respectively.

3.3. Cluster prototype effects

Results from Cox regression analyses using the cluster prototypes on the full cohort and lifetime smoky coal users are provided in tables 2 and 3. The strongest positive associations between pollutant exposure and lung cancer mortality rates were for exposures in the large PAH cluster, both in the full population (HR [95%CI] = 2.4 [2.2, 2.6] per SD) and in lifetime smoky coal users (HR [95%CI] = 1.4 [1.3, 1.5] per SD). In contrast, higher exposures to SO_2 or pollutants from the cluster containing PM, BC, and NO_2 were associated with relatively lower lung cancer mortality.

All exposures in the large PAH cluster showed strong positive associations with lung cancer mortality in single exposure models (figure S5).

Table 2
Hazard ratios for lung cancer mortality for cluster prototypes, estimated using cox regression on data from the full cohort. Baseline rates were stratified by sex and year of birth and the models accounted for confounding by differences in smoking habits and educational status. Hazard ratios were estimated for a one standard deviation change in the estimated prototype exposure.

Cluster[a]	HR [95%CI]
SO2	0.49 [0.46, 0.52]
PM, BC, NO2	0.94 [0.87, 1.01]
NkF	1.20 [1.17, 1.23]
DlP	1.30 [1.24, 1.36]
37 PAHs	2.40 [2.20, 2.62]

[a] Cluster prototype definitions and scaling factors were derived from household air pollutant exposure in controls from nested case-control samples of the full cohort. respectively.

Table 3
Hazard ratios for lung cancer mortality for cluster prototypes, estimated using cox regression on data from lifetime smoky coal users only. Baseline rates were stratified by sex and year of birth and the models accounted for confounding by differences in smoking habits and educational status. Hazard ratios were estimated for a one standard deviation change in the estimated prototype exposure.

Cluster[a]	HR [95%CI]
SO2, PM, BC, NO2	0.80 [0.74, 0.86]
DlP	1.12 [1.06, 1.19]
NkF, ANT	1.21 [1.17, 1.25]
36 PAHs	1.37 [1.28, 1.48]

[a] Cluster prototype definitions and scaling factors were derived from household air pollutant exposure in controls from nested case-control samples of lifetime smoky coal users.

3.4. Timing of exposure

Based on estimated exposure to pollutants from the large PAH cluster, and using the full cohort data, there were 577 cases out of a total of 4,640 (12.4%) that were highly exposed (i.e., higher than average in their sex and age matched controls) only before age 18 years, 173 (3.7%) that were highly exposed only after age 18 years, and 3,303 (71.2%) that were highly exposed during both periods. For control subjects these proportions were 11.8%, 8.8%, and 41.7%, respectively. Estimated HRs of lung cancer mortality for different patterns of exposure are provided in Table 4. The estimated HR of lung cancer mortality for being exposed above the average exposure level both before and after age 18 years is almost identical to that implied by the HRs for being highly exposed either before or after age 18 years.

3.5. Penalized multi-pollutant models

Using a penalized regression modeling framework, ANT and BaC were identified as the strongest risk factors for LC in the nested case-control sample with estimated odds ratios [95%CI] of 3.9 [2.0, 9.2] and 6.7 [0.8, 129] for ANT and BaC respectively (Fig. 1A). Estimated exposure effects were broadly comparable between the model based on data from the full population and that based on data from smoky coal users only, although slightly less pronounced in the latter (Fig. 1B). A graphical comparison of exposure effect estimates with those from our earlier analyses of the retrospective case-control sample is provided in the appendix (figure S6). Results using estimated cumulative exposures after lagging exposure for 10 years were similar to those for unlagged exposures and are not further discussed (appendix figure S7a). Results from the horseshoe regression model fitted to cumulative exposures below age 18 only are also provided in the appendix (figure S7b), but do not point out any specific pollutant risk factor contributing in particular

Table 4
Hazard ratios for lung cancer mortality for high versus low cumulative exposure, assessed either before or after the age of 18, for household air pollutants that are part of the large PAH cluster. Hazard ratios were estimated using cox regression on data from the full cohort. Baseline rates were stratified by sex and year of birth and the models accounted for confounding by differences in smoking habits and educational status.

Exposure pattern[a]	HR [95%CI]
Cumulative exposure from birth to age 18 above risk set average	4.12 [3.76, 4.52]
Cumulative exposure after age 18 above risk set average	1.29 [1.19, 1.40]
	4.14 [3.68, 4.66]
Above risk set average before age 18 only	
Above risk set average after age 18 only	1.30 [1.10, 1.54]
Above risk set average both before and after age 18	5.32 [4.87, 5.80]

[a] Exposure status (either above or below average) was defined with respect to the mean cumulative exposure for controls in each risk-set.

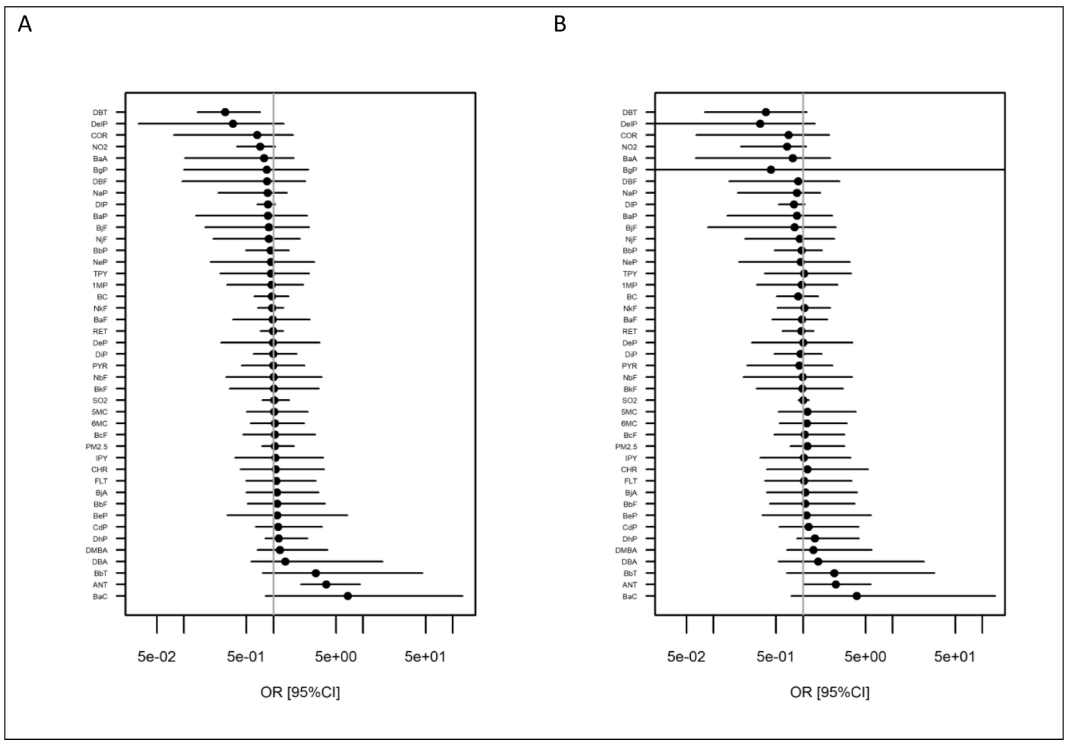

Fig. 1. Odds ratios for death from lung cancer for individual household air pollutants estimated using penalized (horseshoe) conditional logistic regression in nested case-control samples of data from the full cohort (A) and lifetime smoky coal users only (B). Pollutants were ordered by estimated odds ratios for the full cohort sample in both panels. Note: Cases and controls were matched on sex, age, and year of birth and the models accounted for confounding by differences in smoking habits and educational status Scaling factors used to scale exposures before fitting the penalized models were derived using data from the control population for each of the two populations separately (as scaling affects the penalization), but effect estimates from the model fitted to the lifetime smoky coal user population were re-scaled to match those from the full cohort sample.

to the risk observed for early life exposures.

3.6. Non-linear multi-pollutant models

Because the penalized model is likely to struggle with the strong correlations arising from the bimodal distributions for many exposures in this study, we aimed to explore the potential risks associated with exposure to a selected subset of PAH exposures in more detail also in the full cohort data, using multi-pollutant models that allow for non-linear, but smooth, exposure–response relations. For this purpose, we selected ANT and BaC as the strongest risk factors emanating from the penalized multi-pollutant model, BaP based on its known carcinogenicity and because it is often used for risk assessment purposes, and 5MC because it was suggested as a strong risk factor for lung cancer in our earlier case-control study in this population. The full set of results is provided in the appendix (figure S8); results in Fig. 2 are for models that include either BaP or 5MC, because the strong collinearity between their spline functions resulted in strongly diverging exposure–response relations when both were included. All selected PAHs showed a similar, strong positive, exposure–response relation with LC mortality in single-pollutant models, while the exposure–response relation for BaP in a multi-pollutant model that also included ANT and BaC appeared mostly flat, it was clearly positive and strong for 5MC in the comparable model.

4. Discussion

Using data from a large cohort study, we confirmed our earlier finding of a strong exposure–response relation with lung cancer for 5MC, but not for NO_2. We found that exposure to PAHs associated with indoor burning of smoky coal was a strong and significant risk factor for lung cancer mortality. In penalized multi-pollutant models, ANT, rather than 5MC, appeared to be the PAH that was most strongly related to lung cancer, but 5MC still showed a convincing exposure–response relation in regression models where it was included along with ANT and a more restricted set of exposures. We also showed that household air pollutant exposure during childhood was more strongly associated with lung cancer mortality than exposure at later ages.

ANT was identified as the strongest risk factor in both multi-pollutant and non-linear exposure–response models, but there is only limited evidence for its carcinogenicity from experimental animal models (IARC, 2010), and it may therefore be only a proxy for another PAH exposure associated with coal burning. Experimental support for the mutagenicity, carcinogencitiy, and importance of methylated PAHs (including 5MC) in the etiology of lung cancer from smoky coal in Xuanwei is reviewed in Vermeulen *et al.*. (Vermeulen et al., 2019).

Robust identification of single exposure effects proved extremely challenging in the presence of strong between-exposure correlations, even when using state-of-the-art modelling techniques. The fact that we were able to identify only a single, very large, cluster of PAH exposures clearly reflects the strong multi-collinearity among exposures in the present study, which was stronger than in our earlier case-control study where we were able to identify several smaller PAH clusters. Also, whereas NO_2 did not cluster with any other exposure in the case-control study, it did cluster with PM and BC in the present study.

Interpretation of the results from multi-pollutant models can be challenging, especially for correlated exposures, because effect estimates for a single pollutant are conditional on effect estimates for all other pollutants and effects tend to be estimated rather imprecisely. For correlated exposures, estimated HRs$<$1 for some components do not necessarily imply that the exposure has a protective effect, but could also indicate that, in a situation where all exposures occur in combination, risks are lower than might be suggested by linear extrapolation of individual exposure effects. All exposures in the large PAH cluster showed strong positive associations with lung cancer mortality in single

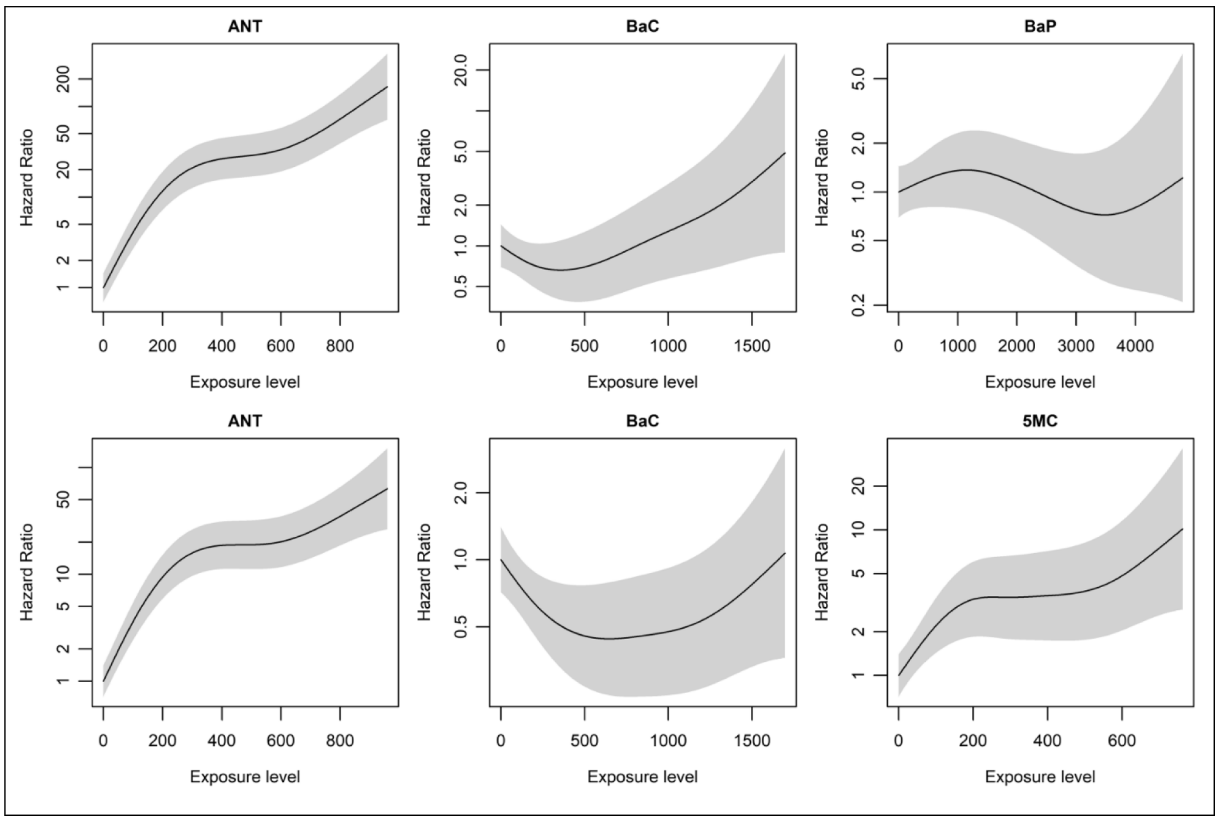

Fig. 2. Non-linear dose–response relations estimated from two different multi-pollutant models that included either anthanthrene (ANT), benzo(a)chrysene (BaC), and benzo(a)pyrene (BaP) (upper panels) or ANT, BaC, and 5–methylchrysene (5MC) (lower panels). Note: Hazard ratios for lung cancer mortality were estimated using cox regression on data from the full cohort, with non-linear exposure–response relations estimated using penalized splines. Baseline rates were stratified by sex and year of birth and the models accounted for confounding by differences in smoking habits and educational status.

exposure models (figure S5).

Average estimated exposure levels in this study were generally in the same range as those estimated for the case-control study (Vermeulen et al., 2019). One important factor is likely the more limited geographical sampling performed in this study, with subjects using smoky coal from a single coal seam only, resulting in a more pronounced bimodal exposure distribution for many PAHs, including 5MC.

The fact that exposure during early life was found to be more strongly related to lung cancer risk than exposure at lower ages fits with other studies that showed the developing lung to be more susceptible to the carcinogenic effects of many traffic-related air pollutants. (Belgrave et al., 2018; Bui et al., 2018; Liu and Kalhan, 2021) Children may be more susceptible to the carcinogens in coal emissions because of increased air intake proportional to lung size (Armstrong et al., 2002) and less efficient carcinogen detoxification. (Bearer, 1995) Importantly, given that childhood is a period of rapid growth and development, it is possible that unrepaired somatically acquired genomic abnormalities in stem cells may carry forward into more daughter cells during adulthood. We were unable to focus on a narrower time-window because reported coal use tended to be stable across most of the childhood, resulting in strong correlations between estimated exposures at different ages. However, most previous research on the effects of early life exposures was limited to studying effects in children or young adults, while the results from our study suggest that significant effects may be found also at much later ages.

Single-pollutant spline-based models showed monotonically increasing exposure–response relations, but this was no longer the case when different, correlated, exposures were included in the same model. In bi-pollutant or higher-order-pollutant models exposure–response curves for BaC and BaP (but not ANT or 5MC) became more variable, even showing a negative dose–response relation for some combinations. Overall, although we found no evidence of strong non-linear exposure–response relations, estimated effects did seem to plateau at high exposures.

We had a large cohort and detailed exposure assessment, with results from earlier studies suggesting that the study was relatively well-powered. However, and as explained earlier, the strong correlations between exposures and the limited geographical sampling meant that power to identify individual exposures was more limited. The use of determinant-based models for exposure assessment may have resulted in shared and complex measurement errors, both between different individuals for the same pollutant, but also between different pollutants for the same individual, which could have affected the results of our multi-pollutant models. (Szpiro and Paciorek, 2013; Evangelopoulos et al., 2021) Our penalized regression models were fitted to data from a nested case-control sample from the full study data for computational reasons, but results from non-penalized models were almost identical when fitted using either Cox regression on the full cohort data or conditional logistic regression on the nested case-control data, confirming the validity of this approach. Estimated effects were reduced in the lifetime smoky coal subpopulation, even after accounting for differences in overall exposure variation. This is likely to be (at least) partly due to a less favourable signal-to-noise ratio in this more highly exposed subpopulation, resulting in attenuated exposure–response relations. However, we cannot fully exclude the possibility that this may also reflect residual confounding by systematic differences between the different communes that used either smoky or smokeless coals.

The diagnosis of lung cancer in this region has traditionally been based on chest X-rays and overall clinical picture, after exclusion of tuberculosis and evidence of other infections. In a minority of cases, a diagnosis was based on cytology or histological examination of tissue. In the first and second follow-ups of the cohort (in 1992 and 1996), 17.6%

of lung cancer diagnoses were cytologically or histologically confirmed. To provide indirect evidence of the validity of a lung cancer diagnosis in the cohort, we compared time from diagnosis to death among cases with and without confirmed diagnoses and found that they were very similar and not statistically significantly different (i.e., 1.21 ± 0.09 years vs. 1.48 ± 0.08 years, respectively). (Lan et al., 2002) In addition, 94.0% of cases died within 3 years of diagnosis. Date of diagnosis was not available from subjects during the third follow-up in 2011, but during the 2000′s in this region all hospitals seeing lung cancer cases were using CT to enhance the diagnosis of lung cancer, so the probability of disease misclassification was further reduced.

Further studies should aim to include subjects that use a wide variety of coals with different chemical composition to allow identification of individual risk factors with more precision. Although cooking and heating practices are rapidly changing in China, (Edwards et al., 2004) a large part of the world is still using solid fuels for cooking and heating. The observation that early life exposures were found to confer most of the risk indicates the urgency for mitigation strategies at a young age. This study adds to the evidence that early life exposures may contribute significantly to air pollution related lung cancer risks later in life.

Funding

The study was supported by the Chinese Academy of Preventive Medicine, Beijing, China, by the Yunnan Province Antiepidemic Station, Kunming, China, and by contract 5D2290NFFX from the US Environmental Protection Agency. This study was also supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health.

CRediT authorship contribution statement

Lützen Portengen: Methodology, Formal analysis, Visualization, Writing - original draft, Writing - review & editing. **George Downward:** Investigation, Data curation, Writing - review & editing. **Bryan A. Bassig:** Investigation, Data curation, Formal analysis, Writing - review & editing. **Batel Blechter:** Data curation, Writing - review & editing. **Wei Hu:** Project administration, Investigation, Data curation, Writing - review & editing. **Jason Y.Y. Wong:** Methodology, Data curation, Formal analysis, Writing - review & editing. **Bofu Ning:** Project administration, Investigation, Data curation, Writing - review & editing. **Mohammad L. Rahman:** Writing - review & editing. **Bu-Tian Ji:** Investigation, Writing - review & editing. **Jihua Li:** Project administration, Investigation, Data curation, Writing - review & editing. **Kaiyun Yang:** Data curation, Writing - review & editing. **H. Dean Hosgood:** Investigation, Writing - review & editing. **Debra T. Silverman:** Conceptualization, Writing - review & editing, Supervision. **Nathaniel Rothman:** Conceptualization, Writing - review & editing, supervision. **Yunchao Huang:** Methodology, Writing - review & editing, Supervision. **Roel Vermeulen:** Conceptualization, Methodology, Formal analysis, Writing - review & editing, Supervision. **Qing Lan:** Conceptualization, Methodology, Investigation, Formal analysis, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We are grateful to the residents of Xuanwei county who participated in this research, and to the late Drs. Robert Chapman and Xingzhou He for their substantial contributions to the study design, field implementation, and data collection for this study. The study was made possible by the outstanding cooperation of many Chinese administrative and public health offices, physicians, and survey workers.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2023.107870.

References

Akaike, H., 1974. A New Look at the Statistical Model Identification. IEEE Trans. Autom. Control 19 (6), 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Andersen, P.K., Gill, R.D., 1982. Cox's Regression Model for Counting Processes: A Large Sample Study. Ann. Stat. 10 (4), 1100–1120. https://doi.org/10.1214/aos/1176345976.

Armstrong, T.W., Zaleski, R.T., Konkel, W.J., Parkerton, T.J., 2002. A tiered approach to assessing children's exposure: a review of methods and data. Toxicol Lett. 127 (1–3), 111–119. https://doi.org/10.1016/s0378-4274(01)00490-8.

Balmes, J.R., 2019. Household air pollution from domestic combustion of solid fuels and health. J. Allergy Clin. Immunol. 143 (6), 1979–1987. https://doi.org/10.1016/J.JACI.2019.04.016.

Barone-Adesi, F., Chapman, R.S., Silverman, D.T., et al., 2012. Risk of lung cancer associated with domestic use of coal in Xuanwei, China: retrospective cohort study. BMJ 345 (aug 292), e5414. https://doi.org/10.1136/bmj.e5414.

Bassig, B.A., Dean Hosgood, H., Shu, X.O., et al., 2020. Ischaemic heart disease and stroke mortality by specific coal type among non-smoking women with substantial indoor air pollution exposure in China. Int. J. Epidemiol. 49 (1), 56–68. https://doi.org/10.1093/ije/dyz158.

Bearer, C.F., 1995. How are children different from adults? Environ Health Perspect. 103 (Suppl 6), 7–12.

Belgrave, D.C.M., Granell, R., Turner, S.W., et al., 2018. Lung function trajectories from pre-school age to adulthood and their associations with early life factors: a retrospective analysis of three population-based birth cohort studies. Lancet Respir Med. 6 (7), 526–534. https://doi.org/10.1016/S2213-2600(18)30099-7.

Box, G.E.P., Cox, D.R., 1964. An Analysis of Transformations. J. Roy. Stat. Soc.: Ser. B (Methodol.) 26 (2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Bui, D.S., Lodge, C.J., Burgess, J.A., et al., 2018. Childhood predictors of lung function trajectories and future COPD risk: a prospective cohort study from the first to the sixth decade of life. Lancet Respir Med. 6 (7), 535–544. https://doi.org/10.1016/S2213-2600(18)30100-0.

Bürkner, P.C., 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. J. Stat. Softw. 80, 1–28. https://doi.org/10.18637/jss.v080.i01.

Carpenter, B., Gelman, A., Hoffman, M.D., et al., 2017. Stan: A Probabilistic Programming Language. J. Stat. Softw. 76, 1–32. https://doi.org/10.18637/jss.v076.i01.

Downward, G.S., Hu, W., Rothman, N., et al., 2014. 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. https://doi.org/10.1021/es504102z.

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

Edwards, R.D., Smith, K.R., Zhang, J., Ma, Y., 2004. Implications of changes in household stoves and fuel use in China. Energy Policy 32 (3), 395–411. https://doi.org/10.1016/S0301-4215(02)00309-9.

Evangelopoulos, D., Katsouyanni, K., Schwartz, J., Walton, H., 2021. Quantifying the short-term effects of air pollution on health in the presence of exposure measurement error: a simulation study of multi-pollutant model results. Environ. Health 20 (1), 94. https://doi.org/10.1186/s12940-021-00757-4.

Gordon, S.B., Bruce, N.G., Grigg, J., et al., 2014. Respiratory risks from household air pollution in low and middle income countries. Lancet Respir. Med. 2 (10), 823–860. https://doi.org/10.1016/S2213-2600(14)70168-7.

Guan, W.J., Zheng, X.Y., Chung, K.F., Zhong, N.S., 2016. Impact of air pollution on the burden of chronic respiratory diseases in China: time for urgent action. Lancet (London, England). 388 (10054), 1939–1951. https://doi.org/10.1016/S0140-6736(16)31597-5.

Hu, W., Downward, G.S., Reiss, B., 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. Tech. 48 (15), 8456–8464. https://doi.org/10.1021/es502201s.

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Some non-heterocyclic polycyclic aromatic hydrocarbons and some related exposures. *IARC Monogr. Eval. Carcinog Risks Hum.* 2010, 92, 1-853.

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 2016. *Outdoor Air Pollution*. International Agency for Research on Cancer.

Lan, Q., Chapman, R.S., Schreinemachers, D.M., Tian, L., He, X., 2002. Household stove improvement and risk of lung cancer in Xuanwei, China. J. Natl Cancer Inst. 94 (11), 826–835. https://doi.org/10.1093/jnci/94.11.826.

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

Lazarevic, N., Knibbs, L.D., Sly, P.D., Barnett, A.G., 2020. Performance of variable and function selection methods for estimating the nonlinear health effects of correlated chemical mixtures: A simulation study. Stat. Med. 39 (27), 3947–3967. https://doi.org/10.1002/sim.8701.

Lee, A., Kinney, P., Chillrud, S., Jack, D., 2015. A Systematic Review of Innate Immunomodulatory Effects of Household Air Pollution Secondary to the Burning of Biomass Fuels. Ann. Glob. Health 81 (3), 368–374. https://doi.org/10.1016/J.AOGH.2015.08.006.

Liu, G.Y., Kalhan, R., 2021. Impaired Respiratory Health and Life Course Transitions From Health to Chronic Lung Disease. Chest 160 (3), 879–889. https://doi.org/10.1016/j.chest.2021.04.009.

Maheswaran, R., Strachan, D.P., Dodgeon, B., Best, N.G., 2002. A population-based case-control study for examining early life influences on geographical variation in adult mortality in England and Wales using stomach cancer and stroke as examples. Int. J. Epidemiol. 31 (2), 375–382. https://doi.org/10.1093/IJE/31.2.375.

Phillips, D.I.W., Osmond, C., Southall, H., Aucott, P., Jones, A., Holgate, S.T., 2018. Evaluating the long-term consequences of air pollution in early life: geographical correlations between coal consumption in 1951/1952 and current mortality in England and Wales. BMJ Open 8 (4). https://doi.org/10.1136/BMJOPEN-2017-018231.

Piironen, J., Vehtari, A., 2017. Sparsity information and regularization in the horseshoe and other shrinkage priors. Electron. J. Statist. 11 (2), 5018–5051. https://doi.org/10.1214/17-EJS1337SI.

Szpiro, A.A., Paciorek, C.J., 2013. Measurement error in two-stage analyses, with application to air pollution epidemiology. Environmetrics 24 (8), 501–517. https://doi.org/10.1002/env.2233.

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

Therneau, T.M., 2022. *A Package for Survival Analysis in R*. https://CRAN.R-project.org/package=survival.

Vermeulen, R., Silverman, D.T., Garshick, E., Vlaanderen, J., Portengen, L., Steenland, K., 2014. Exposure-response estimates for diesel engine exhaust and lung cancer mortality based on data from three occupational cohorts. Environ Health Perspect. 122 (2), 172–177. https://doi.org/10.1289/ehp.1306880.

Vermeulen, R., Downward, G.S., Zhang, J., et al., 2019. Constituents of Household Air Pollution and Risk of Lung Cancer among Never-Smoking Women in Xuanwei and Fuyuan, China. Environ. Health Perspect. 127 (9), 97001. https://doi.org/10.1289/EHP4913.

Vlaanderen, J., Portengen, L., Rothman, N., Lan, Q., Kromhout, H., Vermeulen, R., 2010. Flexible Meta-Regression to assess the shape of the Benzene-Leukemia Exposure-Response curve. Environ. Health Perspect. 118 (4), 526–532. https://doi.org/10.1289/ehp.0901127.

Wong, J.Y.Y., Downward, G.S., Hu, W., et al., 2018. Variation in Lung Cancer Risk by Geologic Coal Deposit Layer: A Case-Control Study of Never-Smoking Women from Xuanwei and Fuyuan, China.