eAppendix

Technical details

Let i index the 4134 miners. Uppercase letters represent random variables and lowercase letters represent realizations of the random variables. Follow-up time was discretized into person-months with age in months indexed by $j = 0$ to 798 and $j = 0$ corresponding to age 18. Miners entered the analysis at age 18 years or age at study entry, whichever occurred last. We administratively censored miners alive at age 90 or on December 31, 2005.

For miner i, V_i represents the time-fixed baseline covariates including race and cumulative radon exposure prior to study entry. Because smoking status was available only as a single binary variable indicating if a miner had ever smoked, smoking was included in V_i as a time-fixed covariate. W_{ij} represents employment status during month j, with W=1 being employed and W=0 being unemployed. All miners enter follow-up employed.Let X_{ij} represent the radon exposure level measured in working level months for miner i during month j. Miners were assumed to be unexposed when not employed in the mine, thus if $W_{ij} = 0$ then $X_{ij} = 0$. Let Y_{ij} = 1 represent death due to lung cancer during month j; $C_{ij} = 1$ represent censoring due to dropout at month j; and $D_{ij} = 1$ represent death due to causes other than lung cancer during month j.

We assume the temporal ordering of the component variables for each month as follows: work status, exposure status, censoring status, death due to other causes, and lung cancer death.

In the equations below, the *i* subscript indexing each miner is suppressed.

The cumulative incidence of lung cancer mortality in the observed data can be written

$$I(j)^{obs}$$

$$= \sum_{k=0}^{j} \sum_{v} \sum_{\bar{w}_j} \sum_{\bar{x}_j} \left\{ P(Y_k = 1 | V = v, \bar{W}_k = \bar{w}_k, \bar{X}_k = \bar{x}_k, S \leq k, \bar{Y}_{k-1} = \bar{D}_k = \bar{C}_k = 0) \right.$$

$$\left. \times \prod_{m=0}^{k} \begin{bmatrix} P(D_m = 0 | V = v, \bar{W}_m = \bar{w}_m, \bar{X}_m = \bar{x}_m, S \leq m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \times \\ P(C_m = 0 | V = v, \bar{W}_m = \bar{w}_m, \bar{X}_m = \bar{x}_m, S \leq m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \times \\ f(X_m = x_m | V = v, \bar{W}_m = \bar{w}_m, \bar{X}_{m-1} = \bar{x}_{m-1}, S \leq m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \times \\ P(W_m = 1 | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{X}_{m-1} = \bar{x}_{m-1}, S \leq m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \times \\ f(V = v) \times \\ P(Y_{m-1} = 0 | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{X}_{m-1} = \bar{x}_{m-1}, S \leq m-1, \bar{Y}_{m-2} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \end{bmatrix} \right\}$$

The cumulative incidence under each intervention was compared to the cumulative incidence under the natural course. The cumulative incidence under the natural course was the cumulative incidence that would have been observed under an intervention that eliminated censoring due to drop-out but did not intervene on exposure.

The cumulative mortality of lung cancer under the natural course is given as:

$$I(j)_{\bar{c}=0}$$

$$= \sum_{k=0}^{j} \sum_{v} \sum_{\bar{w}_j} \sum_{\bar{x}_j} \left\{ P(Y_k = 1 | V = v, \bar{W}_k = \bar{w}_k, \bar{X}_k = \bar{x}_k, S \leq k, \bar{Y}_{k-1} = \bar{D}_k = \bar{C}_k = 0) \right.$$

$$\left. \times \prod_{m=0}^{k} \begin{bmatrix} P(D_m = 0 | V = v, \bar{W}_m = \bar{w}_m, \bar{X}_m = \bar{x}_m, S \leq m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \times \\ 1 \times \\ f(X_m = x_m | V = v, \bar{W}_m = \bar{w}_m, \bar{X}_{m-1} = \bar{x}_{m-1}, S \leq m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \times \\ P(W_m = 1 | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{X}_{m-1} = \bar{x}_{m-1}, S \leq m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \times \\ f(V = v) \times \\ P(Y_{m-1} = 0 | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{X}_{m-1} = \bar{x}_{m-1}, S \leq m-1, \bar{Y}_{m-2} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \end{bmatrix} \right\}$$

To estimate the cumulative incidence under the natural course we fit parametric models for each of the components. Using the 4,134 miners we fit the following models:

1) The probability of remaining at work each month was modeled using logistic regression and was conditional on age, race, calendar year of study entry (year), smoking status (smk), radon dose in the previous month (dose), cumulative number of months worked (mwork), cumulative radon exposure up until the current month for each person that worked the previous month, including any exposure accrued prior to the study (cumrad). The model was run only among at-risk person-time when age was less than 90 years in the current month. Specifically, we fit the model

$$P\left(W_j = 1 | \bar{W}_{j-1} = 1, V = v, \bar{X}_{j-1} = \bar{x}_{j-1}, \bar{Y}_{j-1} = \bar{D}_{j-1} = \bar{C}_{j-1} = 0\right) =$$

$$\text{expit}\Big\{\alpha_0 + \sum_{p=1}^{4} \alpha_p g(AGE) + \alpha_5 RACE + \alpha_6 YEAR + \alpha_7 YEAR^2 + \alpha_8 YEAR^3 + \alpha_9 SMK + \alpha_{10} DOSE_{j-1} + \alpha_{11} DOSE_{j-1}^2 + \alpha_{12} DOSE_{j-1}^3 + \sum_{q=13}^{15} \alpha_q g(MWORK) + \sum_{s=16}^{19} \alpha_s g(CUMRAD)\Big\}$$

where $\text{expit}(.) = \exp(.)/\{1 + \exp(.)\}$ is the anti-logit function, and $g(.)$ represents restricted quadratic splines on age, months worked, and cumulative radon exposure.

2) The monthly density of radon exposure in working level months conditional on age, race, calendar year of study entry, the radon dose from the previous two months, smoking status, and cumulative number of months worked was modeled using linear regression. Specifically, for miners who were at work, we assumed natural log of the monthly radon dose was normally distributed with mean

$$\gamma_0 + \sum_{p=1}^{4} \gamma_p g(AGE) + \gamma_5 RACE + \gamma_6 YEAR + \gamma_7 YEAR^2 + \gamma_8 YEAR^3 + \gamma_9 DOSE_{j-1} + \gamma_{10} DOSE_{j-2} + \gamma_{11} SMK + \gamma_{12} MWORK + \gamma_{13} MWORK^2 + \gamma_{14} MWORK^3$$

and variance σ^2, estimated by the mean squared error.

3) The probability of competing death was predicted on the basis of age, race, calendar year of study entry, cumulative number of months worked, past time-varying work

status, smoking status, and cumulative radon exposure. Specifically, we fit the model:

$$P\left(D_j = 1 | V = v, \overline{W}_j = \overline{w}_j, \overline{X}_j = \overline{x}_j, \overline{Y}_{j-1} = \overline{D}_{j-1} = \overline{C}_{j-1} = 0\right) =$$

$$\text{expit}\left\{\delta_0 + \sum_{p=1}^{4} \delta_p g(AGE) + \delta_5 RACE + \delta_6 YEAR + \delta_7 YEAR^2 + \delta_8 YEAR^3 + \delta_9 MWORK + \delta_{10} MWORK^2 + \delta_{11} MWORK^3 + \delta_{12} WORK_j + \delta_{13} WORK_{j-1} + \delta_{14} WORK_{j-2} + \delta_{15} SMK + \sum_{q=16}^{19} \delta_q g(CUMRAD)\right\}$$

4) The probability of lung cancer mortality was modeled conditional on age, race, calendar year of study entry, cumulative number of months worked, past time-varying work status, smoking status, and cumulative radon exposure. Specifically, we fit the model:

$$P\left(Y_j = 1 | V = v, \overline{W}_j = \overline{w}_j, \overline{X}_j = \overline{x}_j, \overline{Y}_{j-1} = \overline{D}_{j-1} = \overline{C}_{j-1} = 0\right) =$$

$$\text{expit}\left\{\beta_0 + \sum_{p=1}^{4} \beta_p g(AGE)_1 + \beta_5 RACE + \beta_6 YEAR + \beta_7 YEAR^2 + \beta_8 YEAR^3 + \beta_9 MWORK + \beta_{10} MWORK^2 + \beta_{11} MWORK^3 + \beta_{12} WORK_j + \beta_{13} WORK_{j-1} + \beta_{14} WORK_{j-2} + \beta_{15} SMK + \sum_{q=16}^{19} \beta_q g(CUMRAD)\right\}$$

As a sensitivity analysis, we included an interaction between cumulative radon exposure and exposure rate in model (3b) below for the probability of lung cancer death. Because results were unchanged, we used the simpler model (3) shown above.

3b)

$$P\left(Y_j = 1 | V = v, \bar{W}_j = \bar{w}_j, \bar{X}_j = \bar{x}_j, \bar{Y}_{j-1} = \bar{D}_{j-1} = \bar{C}_{j-1} = 0\right) =$$

$$\begin{aligned}\text{expit}\Big\{\beta_0 + \sum_{p=1}^{4} \beta_p g(AGE)_1 + \beta_5 RACE + \beta_6 YEAR + \beta_7 YEAR^2 + \beta_8 YEAR^3 + \beta_9 MWORK \\ + \beta_{10} MWORK^2 + \beta_{11} MWORK^3 + \beta_{12} WORK_j + \beta_{13} WORK_{j-1} \\ + \beta_{14} WORK_{j-2} + \beta_{15} SMK + \sum_{q=16}^{19} \beta_q g(CUMRAD) \\ + \sum_{q=20}^{23} \beta_q g(CUMRAD) \times EXPRATE\Big\},\end{aligned}$$

where EXPRATE was defined as the cumulative radon exposure divided by the total time exposed to radon.

To estimate the lung cancer mortality under various interventions, we drew a large Monte Carlo sample of miners (N=50,000) with replacement from the original data. For each sampled miner, we predicted work status, exposure, and outcomes in each month under each intervention scenario. Each intervention takes the form, "if the miner is at work, and if predicted monthly exposure exceeds the intervention exposure limit set exposure to the intervention exposure limit; if the miner is at work and the predicted monthly exposure is below the intervention exposure limit, do not intervene on that miner at that time; if the miner is not at work, set exposure to 0." All interventions allowed no censoring due to dropout.

Accordingly, the g-formula for cumulative lung cancer mortality under the intervention scenarios limiting radon exposure was

$$I(j)_{\bar{x}_j,\bar{c}=0}$$

$$= \sum_{k=0}^{j} \sum_{v} \sum_{\bar{w}_j} \sum_{\bar{x}_j} \left\{ P(Y_k = 1 | V = v, \bar{W}_k = \bar{w}_k, \bar{x}_k, S \leq k, \bar{Y}_{k-1} = \bar{D}_k = \bar{C}_k = 0) \right.$$

$$\left. \times \prod_{m=0}^{k} \left[\begin{array}{c} P(D_m = 0 | V = v, \bar{W}_m = \bar{w}_m, \bar{x}_m, S \leq m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \times \\ 1 \times \\ f(X_m = x_m | X_m^* = x_m^*, V = v, \bar{W}_m = \bar{w}_m, \bar{X}_{m-1} = \bar{x}_{m-1}, S \leq m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \times \\ f(X_m^* = x_m^* | V = v, \bar{W}_m = \bar{w}_m, \bar{X}_{m-1} = \bar{x}_{m-1}, S \leq m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \times \\ P(W_m = 1 | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{x}_{m-1}, S \leq m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \times \\ f(V = v) \times \\ P(Y_{m-1} = 0 | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{x}_{m-1}, S \leq m-1, \bar{Y}_{m-2} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0) \end{array} \right] \right\}$$

where $\bar{x}_j$ describes the exposure history up to time *j* that would have been observed for a miner had he been subject to that intervention scenario. X_m^* represents the predicted exposure for month *m* in the absence of an exposure limit in month *m.* X_m is the exposure for month *m* after the exposure limit is applied to the predicted exposure X_m^*. See Taubman 2009 and Young 2014 for further details on estimating the effects of interventions that depend on the natural value of exposure.