APPENDIX

Notation and Definitions

Consider a longitudinal study of n subjects, with each subject's observations assumed to be independent and identically distributed. Time is indexed by k, for $k = 1, 2, 3, \cdots, 10$. V is a vector for baseline covariates. A_k is a measure of $PM_{2.5}$ levels participants were exposed to at time k. L_k is a health status variable, here a continuous comprehensive risk score variable based on health insurance claims. Y_k is defined as continuous outcome of interest (here height standardized FEV_1 or FVC in separate analyses) at time k. C_k is an indicator for censoring due to termination of employment or death at time k. The order assumed at each time k is V_k, L_k, A_k,C_k, Y_k. The relationships between exposure A, time-varying covariate L, censoring C and lung function parameters Y are also depicted in a directed acyclic graph in Figure 1 of the main manuscript. Histories for time varying covariates are denoted with overbars such that exposure history through time k is denoted by $\bar{A}_k = \{A_1, A_2, A_3, \cdots, A_k\}$.

The parametric g-formula

The expected outcome $E[Y_k]$ for lung function parameters FEV_1 and FVC at time $k = 10$ is given below as a function of the joint density of the outcome conditional on exposures and covariates and the joint distribution of exposure and covariates (Taubman et al. 2009).

$$\sum_{v}\sum_{\bar{l}_{10}}\sum_{\bar{\alpha}_{10}}\left\{\begin{array}{l} E[Y_k|V=v,\bar{L}_k=\bar{l}_k,\bar{A}_k=\bar{\alpha}_k,\bar{C}_k=0,S\leq k] \\ \times\prod_{j=1}^{k}\left[\begin{array}{l} Pr[C_j=0|V=v,\bar{L}_j=\bar{l}_j,\bar{A}_j=\bar{\alpha}_j,\bar{C}_{j-1}=0]\times \\ f[A_j=\alpha_j|V=v,\bar{L}_j=\bar{l}_j,\bar{A}_{j-1}=\bar{\alpha}_{j-1},\bar{C}_{j-1}=0]\times \\ f[L_j=l_j|V=v,\bar{L}_{j-1}=\bar{l}_{j-1},\bar{A}_{j-1}=\bar{\alpha}_{j-1},barC_{j-1}=0,]\times \\ f(V=v)\times \end{array}\right] \end{array}\right\}$$

Under a simulated interventions where A_k in intervened on (in our case deterministically) the above quantity reduces to:

$$\sum_{v}\sum_{\bar{l}_{10}}\left\{\begin{array}{l} E[Y_k|V=v,\bar{L}_k=\bar{l}_k,\bar{A}_k=\bar{\alpha}_k,\bar{C}_k=0,S\leq k] \\ \times\prod_{j=1}^{k}\left[\begin{array}{l} Pr[C_j=0|V=v,\bar{L}_j=\bar{l}_j,\bar{A}_j=\bar{\alpha}_j,\bar{C}_{j-1}=0]\times \\ f[L_j=l_j|V=v,\bar{L}_{j-1}=\bar{l}_{j-1},\bar{A}_{j-1}=\bar{\alpha}_{j-1},barC_{j-1}=0,]\times \\ f(V=v)\times \end{array}\right] \end{array}\right\}$$

Similarly in the sensitivity analyses for simulated interventions where no participant is assumed to be censored:

$$\sum_{v}\sum_{\bar{l}_{10}}\left\{\begin{array}{l} E[Y_k|V=v,\bar{L}_k=\bar{l}_k,\bar{A}_k=\bar{\alpha}_k,\bar{C}_k=0,S\leq k] \\ \times\prod_{j=1}^{k}\left[\begin{array}{l} f[L_j=l_j|V=v,\bar{L}_{j-1}=\bar{l}_{j-1},\bar{A}_{j-1}=\bar{\alpha}_{j-1},barC_{j-1}=0,]\times \\ f(V=v)\times \end{array}\right] \end{array}\right\}$$

This quantity is the extension of standardization for time-varying exposures (Taubman et al. 2009). The expression above is a sum over all possible covariate and exposure histories, which cannot be computed non-parametrically in high dimensional data such as the dataset in question. We instead rely on the use of parametric models and a Monte Carlo simulation to approximate the g-formula.

The parametric models give us estimates for the individual probabilities in the above expression. Models were fit pooled on the person-year level. For each intervention considered, exposure and covariate histories are generated for a Monte Carlo sample based on the observed distribution of baseline covariates, and using the predicted probabilities and densities from the parametric models fitted to simulate values over time. The expectation for the outcome at time k is then estimated in each simulated dataset.

The process is performed in a SAS macro in the following steps:

Parametric models

We fit parametric models to estimate the above probabilities and densities. Baseline covariates V were chosen *a priori* and entered in all models as follows: cubic polynomials for age and calendar year with user defined knots, indicator variables for facility type (smelter, fabrication or refinery), facility lovation (8 different locations), job grade (above or blow median pay), non-white race, smoking status (ever, never or missing) and a continuous variable for cumulative $PM_{2.5}$ exposure

accrued prior to beginning of follow up. The degree of paramerization and knot placement for the age variable was varied and ultimately chosen based on model fit. This vector function is collectively portrayed as $g_1(v)$ in the following models.

The parametric models fit were as follows:

a. A linear model for the level of annual daily average $PM_{2.5}$ exposure at time k,

$E[A_k|V = v, \bar{L}_k = \bar{l}_k, \bar{A}_{k-1} = \bar{\alpha}_{k-1}, \bar{C}_{k-1} = 0] = \beta_0 + \beta_1^{'} g_1(v) + \beta_2^{'} g_2(\bar{l}_k) + \beta_3^{'} g_3(\bar{\alpha}_{k-1})$, where $\beta_1^{'}$ is a vector of parameter coefficients for all baseline covariates as listed above, β_2 is a vector of parameter coefficients for the terms of a function of risk score history $g_2(\bar{l}_k)$, specifically a cubic spline term for risk score at time k. Finally $\beta_3^{'}$ is a vector of parameter coefficients for the terms of a function of $PM_{2.5}$ exposure history $g_3(\bar{\alpha}_{k-1})$, specifically cubic spline term for $PM_{2.5}$ exposure levels at time $k-1$.

b. A linear model for the level of risk score at time k,

$E(L_k = l(k)|V = v, \bar{L}_{k-1} = \bar{l}_{k-1}, \bar{A}_{k-1} = \bar{\alpha}_{k-1}, \bar{C}_{k-1} = 0) = \delta_0 + \delta_1^{'} g_1(v) + \delta_2^{'} g_2(\bar{l}_{k-1}) + \delta_3^{'} g_3(\bar{\alpha}_{k-1})$, where $\delta_1^{'}$ is a vector of parameter coefficients for all baseline covariates, and $\delta_2^{'}$ is a vector of parameter coefficients for the terms of a function of risk score history $g_2(\bar{l}_{k-1})$, specifically a cubic spline term for risk score at time $k-1$, and $\delta_3^{'}$ is a vector of parameter coefficients for the terms of a function of $PM_{2.5}$ exposure history $g_3(\bar{\alpha}_{k-1})$, specifically a cubic spline term for $PM_{2.5}$ exposure levels at time $k-1$.

c. A logistic model for the probability of censoring at time k,

$logit(C_k|V = v, \bar{A}_k = \bar{\alpha}_k, \bar{L}_k = \bar{l}_k, \bar{C}_{k-1} = 0) = \theta_0 + \theta_1^{'} g_1(v) + \theta_2^{'} g_2(\bar{l}_k) + \theta_3^{'} g_3(\bar{\alpha}_k)$, where $\theta_1^{'}$ is a vector of parameter coefficients for all baseline covariates, $\theta_2^{'}$ is a vector of parameter coefficients for a function of location history $g_2(\bar{l}_k)$ as in (a), $\theta_3^{'}$ is a vector of parameter coefficients for a function of exposure history $g_3(\bar{\alpha}_k)$, specifically a cubic spline term, for the exposure levels at time k plus a cubic spline for cumulative $PM_{2.5}$ exposure up to time $k-1$.

d. Finally, a linear model for height standardized FEV_1 or FVC at time k,

$E(Y_k|V = v, \bar{A}_k = \bar{\alpha}_k, \bar{L}_k = \bar{l}_k, \bar{C}_k = 0) = \psi_0 + \psi_1^{'} g_1(v) + \psi_2^{'} g_2(\bar{l}_k) + \psi_3^{'} g_3(\bar{\alpha}_k)$, where $\psi_1^{'}$ is a vector of pa-

rameter coefficients for all baseline covariates, and $\psi_2^{'}$ and $\psi_3^{'}$, are vectors of parameter coefficients for the terms of risk score, and exposure histories respectively, as described in (c).

Simulation

We next draw a Monte Carlo sample (N=50,000). Values for baseline covariates V and baseline values for exposure and work status are set as observed at time $k = 1$. By definition the following is also true: $C_{k=1} = 0$.

Under the natural course (no intervention), for all $k > 1$, A_k, L_k, C_k and Y_k are predicted using the parameters from the fitted models described above and the simulated values of all covariate and exposure histories up to that age. Values for linear terms (in this case the exposure term A_k, risk score L_k and outcome Y_k), are determined using the model predicted values plus a random error term, while values for discrete variables (here censoring C_k) are determined using the model predicted probability and a randomly drawn uniform value. Values of one are assigned if the predicted probability is below this value and zero otherwise.

Interventions

Under an intervention setting a deterministic value for the exposure, the values of A_k are set according to the intervention defined value. Under an intervention setting an exposure limit, the values of A_k, are intervened on if the model based simulated value is above the theoretical limit by replacing the predicted exposure with the value of the theoretical limit. The simulation continues using the intervened values for the prediction of subsequent exposure and covariate values. The simulation continues until time $k = 10$ or until C_k is simulated to be 1. C_k is set to zero at all times in the sensitivity analyses assuming no one is censored, in which case the simulation continues to time $k = 10$ for everyone.

The estimated outcome is averaged over the simulated population, summed over all covariate histories and weighted by the frequency of covariate histories, giving the g-formula estimate described above. Differences in height standardized FEV_1 or FVC are estimated for each intervention by

comparing the outcome under each interventions to the outcome under no intervention (natural course).

Confidence Intervals

For measures of stability we construct 95% Confidence Intervals (CI) by repeating the entire process in 200 bootstraps (sampled from the observed data with replacement). Each bootstrapped sample was of size=6485. We use the standard deviation of the bootstrapped difference estimates as an estimate of the standard error to construct CIs.