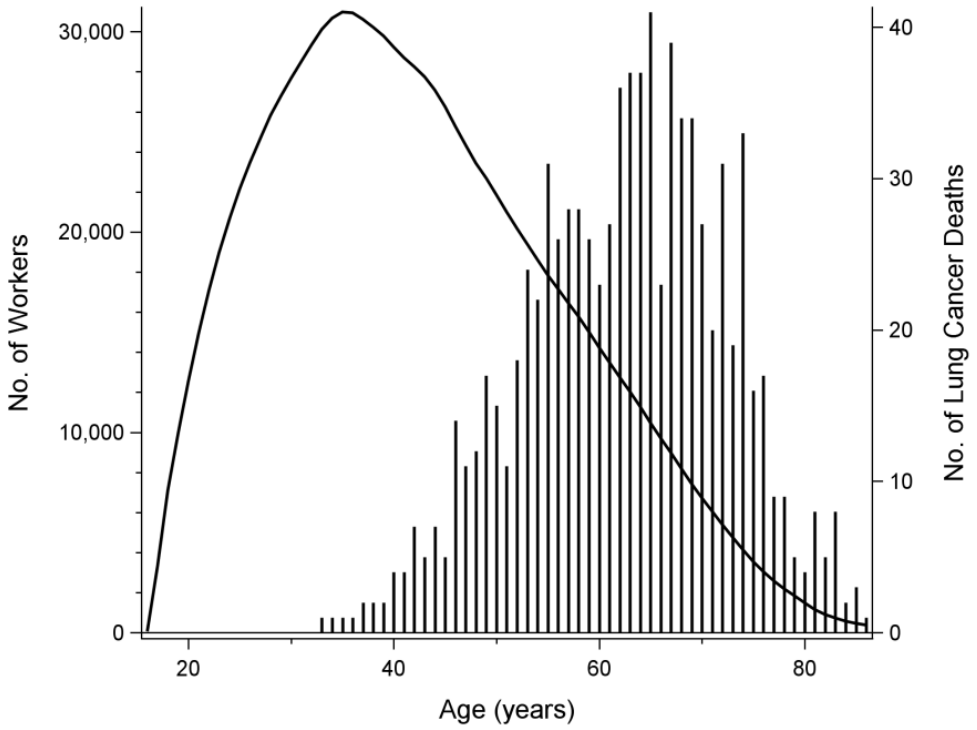
Online Supplementary File



Supplementary Figure S1. Number of workers under follow-up (curved line) and number of lung cancer deaths (vertical lines) by age among 38,560 autoworkers during the 1,122,160 person-years of follow-up between January 1, 1941 and December 31, 1994.

Supplementary Table S1. Exposure characteristics over person-years in the observed data and simulated interventions[a] in the UAW-GM cohort. [MWF=metalworking fluid; IQR=interquartile range]

	Observed			Simulated Natural Course			Synthetic Fluid Intervention[a]		
	%	Median	IQR	%	Median	IQR	%	Median	IQR
Employed[b]	76.7			63.8			64.2		
MWF[bc]									
Straight	21.7	0.08	0.04–0.43	13.9	0.09	0.03–0.35	12.5	0.11	0.03–0.40
Soluble	47.4	0.48	0.20–0.88	29.3	0.26	0.12–0.65	28.8	0.27	0.12–0.66
Synthetic	10.9	0.09	0.01–0.13	10.8	0.02	0.01–0.08	0.0	·	·
Biocide									
Ever Exposed	25.4			32.0			32.0		

[a]Interventions intervened on the synthetic fluid exposure and fixed exposure to always unexposed, i.e., a ban.
[b]Lagged 15 years
[c]Percent of person-years exposed and annual exposure (mg/m^3) distribution among exposed

Supplementary Appendix S1

An assumption of the parametric g-formula is correct model specification. However, baseline variables did not require parametric models as these values were drawn from the empirical distribution. While we do intervene upon the exposure variables in the simulated datasets, we do not intervene on all the exposures all of the time. For some interventions, exposure variables are treated simply as time-varying covariates. Thus, we assumed correct specifications of the parametric models for employment status, death due to a competing risk, death due to lung cancer, and for biocide and metalworking fluid (MWF) exposures, including both annual and cumulative models for each of straight, soluble, and synthetic MWF. As an informal check of this assumption, we compared the cumulative lung cancer mortality in the observed data to the cumulative lung cancer mortality in the simulated natural course. Specification of the models described below was done with the goal of minimizing the difference between the observed and simulated natural course cumulative lung cancer mortality. The observed cumulative lung cancer mortality was 7.201%. The final models, as described below, achieved a simulated a natural course cumulative lung cancer mortality of 7.202%.

Baseline covariates were entered in all models as follows: indicator variables for sex (male, female) and race (white, African American), and plant (1, 2, 3); linear term for year of hire; categorical variable for calendar year with cut points every five years starting with 1950 through 1990; categorical variable for age with a level per decade starting with the 30s. Calendar year and age were simulated as time-varying covariates that increased by one year for each subject for each person-year record after baseline. Age was the time-scale of interest, indexed $a = 16$ to 86. All models were conditional on prior survival.

The parametric models were fit as follows:

1. A logistic model for the probability of employment termination status at age a, restricted to records where prior employment termination status = 0 (not terminated). Because of the 15-year lag, the first 15 records were set to 0. As an independent variable for succeeding covariate models, the prior value of employment termination status was included as an indicator variable.

2. A two-stage process for the level of annual average daily straight metalworking fluid (MWF) exposure at age a. First, a logistic model was used for the probability of straight MWF exposure being greater than 0. Second, for records in which exposure was predicted to be greater than 0, a linear model predicted the natural log of the straight MWF exposure level. In records in which employment termination status = 1 (terminated), exposure was set to 0. Because of the 15-year lag, the first 15 records were set to 0 exposure. As an independent variable for succeeding covariate models, the prior two values of straight MWF exposure were included as categorical variables with cut points at the quintiles of annual exposure among all observed person-time: 0.0295, 0.0569, 0.164, and 0.5395 mg/m^3.

3. A model accumulating annual average daily straight MWF exposure to predict cumulative straight MWF exposure at age a. Because of the 15-year lag, the first 15 records were set to 0 exposure. Values of cumulative straight MWF exposure were not used as independent variables in covariate models. These were only used as independent variables in the models for lung cancer death and non-lung cancer death, specifically the prior two values of cumulative straight MWF exposure were included as restricted cubic splines with knots at the quintiles of cumulative exposure among observed lung cancer deaths: 0.0871, 0.3401, 0.9497, and 2.8895 mg/m^3-years.

4. A two-stage process for the level of annual average daily soluble metalworking fluid (MWF) exposure at age a. First, a logistic model was used for the probability of soluble MWF exposure being greater than 0. Second, for records in which exposure was predicted to be greater than 0, a linear model predicted the natural log of the soluble MWF exposure level. In records in which employment termination status = 1 (terminated), exposure was set to 0. Because of the 15-year lag, the first 15 records were set to 0 exposure. As an independent variable for succeeding covariate models, the prior two values of soluble MWF exposure were included as categorical

variables with cut points at the quartiles of annual exposure among all observed person-time: 0.2047, 0.4767, and 0.8841 mg/m^3.

5. A model accumulating annual average daily soluble MWF exposure to predict cumulative soluble MWF exposure at age *a*. Because of the 15-year lag, the first 15 records were set to 0 exposure. Values of cumulative soluble MWF exposure were not used as independent variables in covariate models. These were only used as independent variables in the models for lung cancer death and non-lung cancer death, specifically the prior two values of cumulative soluble MWF exposure were included as restricted cubic splines with knots at the quintiles of cumulative exposure among observed lung cancer deaths: 1.4934, 3.5005, 6.9223, and 13.5493 mg/m^3-years.

6. A two-stage process for the level of annual average daily synthetic metalworking fluid (MWF) exposure at age *a*. First, a logistic model was used for the probability of synthetic MWF exposure being greater than 0. Second, for records in which exposure was predicted to be greater than 0, a linear model predicted the natural log of the synthetic MWF exposure level. In records in which employment termination status = 1 (terminated), exposure was set to 0. Because of the 15-year lag, the first 15 records were set to 0 exposure. As an independent variable for succeeding covariate models, the prior two values of synthetic MWF exposure were included as categorical variables with cut points at the quintiles of annual exposure among all observed person-time: 0.0143, 0.0286, 0.0878, and 0.1831 mg/m^3.

7. A model accumulating annual average daily synthetic MWF exposure to predict cumulative synthetic MWF exposure at age *a*. Because of the 15-year lag, the first 15 records were set to 0 exposure. Values of cumulative synthetic MWF exposure were not used as independent variables in covariate models. These were only used as independent variables in the models for lung cancer death and non-lung cancer death, specifically the prior two values of cumulative synthetic MWF exposure were included as restricted cubic splines with knots at the quartiles of cumulative exposure among observed lung cancer deaths: 0.1347, 0.5216, and 1.3826 mg/m^3-years.

8. A logistic model for the probability of biocide exposure at age *a*. Restricted to records where prior biocide exposure = 0 (never exposed). Values of biocide exposure were not used as independent variables in covariate models. These were only used as independent variables in the models for lung cancer death and non-lung cancer death, specifically the prior value of biocide exposure was included as an indicator variable (never or ever exposed).

9. A logistic model for the probability of death due to causes other than lung cancer at age *a*.

10. A logistic model for the probability of lung cancer death at age *a*.