eAppendix

Parametric g-formula algorithm – algorithmic details

The generalize computation algorithm formula (g-formula) is a way to describe the relationships between exposure, confounders, and potential outcomes (e.g. the expected lung cancer mortality at age 80 under no exposure). The g-formula links to the observed data via the causal identification assumptions. The relationships among the observed data can be modeled using parametric modeling, in which case the method is referred to as the parametric g-formula. This method is highly parametric because it requires estimating nuisance parameters (e.g. log-odds ratios from the parametric models) in order to estimate the target parameter (e.g. the age-specific risk of the outcome under no exposure). The parameters of the parametric g-formula can be estimated using the combination of a set of pooled logistic models for the joint distribution of the data, followed by a Monte Carlo algorithm.[1,50] The modeling step can be performed using separate models for exposure, all time-varying covariates, and all outcomes of interest (see below). The Monte Carlo step uses parameter estimates from the modeling step to simulate the target population under one or more intervention distributions for exposure. The output of these simulations is a set of discrete hazard estimates for each outcome yearly for each subject under each intervention. Using a modified Kaplan-Meier algorithm, these discrete rate estimates are combined to estimate the risk for the outcomes of interest under each intervention. The risk difference comparing two interventions at a given age is simply the difference between the risk functions. Interval estimates for the risk functions and the risk difference estimates are obtained using a non-parametric bootstrap, whereby all steps of the parametric g-formula are repeated on random samples (with replacement) of individuals from the original data. We describe each step in more detail below.

Parametric modeling

Using a person-period dataset where each observation represents a person-year, we fit the following models: a) a logistic model to estimate the log-odds of whether or not the individual was exposed at all during the year, if at work (Chinese, US diatomaceous studies only, in the other cohorts subjects were exposed until they left work); b) a linear model to estimate the log-annual-exposure rate (a and b are referred to jointly as 'the exposure model'); c) a logistic model to estimate the log-odds of leaving employment in a given year, if at work (the 'employment model'); d) a logistic model to estimate the log-odds of dying from lung cancer in a given year (the 'lung cancer model'); e) a logistic model to estimate the log-odds of dying from a cause other than lung cancer in a given year (the 'other outcome model'). The variables used in each model are given in Supplemental Table 4. Models were chosen from a list of candidate models based on Aikiake's Information Criterion (AIC) and comparisons between the observed data and the so-called 'natural course' intervention. The natural course intervention refers to an intervention that attempts to emulate the existing data by modeling the exposure mechanism in addition to confounders and outcomes. We included all potential confounders and model selection consisted of examining model fit under different functional forms for variables. We performed an initial models selection based on AIC, which yielded a small set of potential model forms for each modeled variable. We selected our final set of models from this set based on plots to assess agreement between the between the natural course and the observed data, with respect to risk.

Monte Carlo algorithm

We fully describe the Monte Carlo algorithm only for the natural course, and variations of the algorithm under interventions are considered below. We previously described a simpler form of this algorithm in detail,[1] and we described an occupational implementation in a cohort of copper smelters.[2] The Monte Carlo algorithm starts out using a large (M =1,000,000) random sample, taken with replacement from the study population (N = 65,999) individuals. Because of large difference between cohort sizes, it requires a very large random sample to minimize simulation error in the smaller cohorts. For computational efficiency, we sampled equally from each cohort and recorded sampling fractions to use as weights in subsequent estimation procedures. The values of all study variables, except exposure, were retained for the person-year in which the individual entered the study (referred to as the baseline variables). The sampled individuals are referred to as members of a "pseudo-population."

For each member of the pseudo-population, the algorithm to predict the outcomes under each intervention proceeds from age, in years, at study entry (time T_i=*1* for $i = 1, \ldots, M$ pseudo-individuals) until the age at death or the age at which follow-up for the pseudo-individual would have ended (time $T_i = K_i$). Starting at time T_i=*1* for the first individual (i=*1*), we calculate the predicted probability of whether the pseudo-individual leaves work using values of their baseline covariates and the model coefficients for the employment model. We then draw a value from a Bernoulli distribution with this probability to simulate termination from work. If the pseudo-individual leaves employment then exposure at time $T_i=1$ is set to zero. If the pseudo-individual remains employed, then we predict their exposure. For diatomaceous earth workers and the 4 Chinese cohorts, some individuals were unexposed at work, and we predicted whether or not a pseudo-individual was exposed for these four cohorts using their values of the baseline covariates, and the coefficients from the logistic (yes/no) exposure model. Among those predicted to be exposed, we simulated log-annual-exposure for each pseudo-individual using their values of the baseline covariates, and the coefficients from the exposure model. For the natural course, log-annual-exposure is taken as a draw from a normal distribution with the mean set at the prediction and the variance set at the estimated error variance from the linear model for log-exposure. For unexposed members of the diatomaceous earth, or any of the 3 Chinese cohorts, we set exposure to a small, non-zero value if the individual was predicted to have been unexposed while at work. This constant was set to a value for annual exposure determined by the smallest observed annual exposures by country (exp(-11) and exp(-7) in the diatomaceous earth and Chinese cohorts, respectively). Given that these exposures are several orders of magnitude below any intervention value, the results are not expected to be sensitive to these values because they are not used in modeling quantitative exposure, and outcomes are modeled using untransformed exposures.

Using baseline covariates and the new predicted employment and exposure for time $T_i=1$ and the lung cancer model, we then predict whether or not the pseudo-individual dies from lung cancer during the first year of employment. If the individual does not die from lung cancer, we use the predicted employment and exposure for time $T_i=1$ and the other outcome model to generate a predicted probability for whether or not the pseudo-individual dies from another cause during the year.

If the individual does not die or reach the end up follow-up, then we set $T_i=T_i+1$ and repeat the algorithm. The algorithm is repeated until the pseudo-individual dies or reaches the end of follow-up. Once a pseudo-individual dies or reaches the end of follow-up, we set $i=i+1$ and repeat the algorithm for the next pseudo-individual, and so on for all individuals in the pseudo-population. Once the algorithm has moved through the entire pseudo-population, we have a "pseudo-cohort" of person-period data that is a realization of the joint-probability distribution implied by the parametric g-formula. For large *M* (and provided that the causal identification and model specification assumptions are met), the empirical distribution of the pseudo-cohort estimates the joint probability distribution of the data we would expect under intervention (or no intervention, in the case of the natural course). Thus, the risk we would expect in the target population, had an intervention occurred, is the sample risk in the pseudo-cohort. Note that this algorithm yields the risk on the time-scale with an origin at the start of follow-up. We estimate risk on the age time-scale using an approach described below.

Interventions other than the natural course

The algorithm described above yields the mortality under no intervention on exposure. To estimate the mortality under interventions on occupational silica exposure, we modify the sampling from the exposure model to draw log-exposures from a truncated normal distribution, with an upper bound at the proposed limit. For example, for an occupational exposure limit of 10 μg/m^3, we would first draw a value from the log-exposure distribution implied by the natural course. If that value is above log(10), then we repeat the sampling until we draw a value at or below log(10).

Cumulative incidence, risk difference estimation

We estimate risk from age 16 to 90 for each cause of death and intervention using a modified version of the Kaplan-Meier estimator.[2,3] We modify previous implementations slightly by sampling from each cohort with equal probability and using sampling weights to account for this differential sampling when estimating the risk. This approach was needed to reduce simulation error in estimating cohort and country specific mortality, given computational constraints.

Sensitivity analysis

We assessed sensitivity to our particular choice of functional form between exposure and mortality in the models for lung cancer mortality and all other cause mortality by re-estimating the parameters of the g-formula using modified outcome models that included spline terms for total exposure from 1 to 14 years prior (knots at 0.0, 0.4, 1.3, 20.0), and cumulative exposure lagged 15 years (knots at 0.0, 0.0, 1.4, 23.0),. Both splines were restricted cubic splines with basis functions with 3 degrees of freedom. Results are given in eTable 8.

For the Chinese cohorts, we also estimated risk with the g-formula while operationalizing silicosis as an additional time-varying-confounder. We assumed an identical model form for silicosis as for mortality (see eTable 1), and we included silicosis at baseline and ever-silicosis (time-varying) as indicator variables in models for exposure and mortality. Results are given in eTable 5.

eTables

eTable 1: Model forms used for the parametric g-formula analysis

Dependent var.	Studies	Model	Terms[a]
Exposure (unexposed at work=0 vs. exposed at work=1)	Chinese, US diatomaceous	Logistic model	agec agework_sp1 datec datework_sp1 datework_sp2 sex[a] race[b] explag1 cye1_3_1 cye1_3_2
Exposure (quantitative)	Non-Chinese	Linear model for log-exposure	agec agework_sp1 datec datework_sp1 datework_sp2 sex[b] race[b] explag1 exp_2_15 cumexplag15 cumyrsexp cumyrsexp*cumyrsexp
Exposure (quantitative)	Chinese	Linear model for log-exposure	agec agework_sp1 datec datework_sp1 datework_sp2 sex[b] race[b] xl1cat_5_2 xl1cat_5_3 xl1cat_5_4 sil2_4_1 sil2_4_2 sil2_4_3 cumyrsexp cumyrsexp*cumyrsexp
Employment (leaving employment=1 vs. staying employed=0)	All	Logistic model	agec agesq agecu datec datesq datecu sex[a] race[b] cumexp
Lung cancer mortality	All	Logistic model	agec agedeath_sp1 agedeath_sp2 datec datedeath_sp1 datedeath_sp2 sex[b] race[b] atwork_2yrs exp_2_15 cumexplag15 cumyrsexp cumyrsexp*cumyrsexp
All other causes of death	All	Logistic model	agec agedeath_sp1 agedeath_sp2 datec datedeath_sp1 datedeath_sp2 sex[b] race[b] atwork_2yrs exp_2_15 cumexplag15 cumyrsexp cumyrsexp*cumyrsexp

[a] Key for variable names: agec (attained age, centered), agedeath_sp1-2 and agework_sp1_2 (restricted, cubic spline on agec; knots varied for Chinese vs. other cohorts), datework_sp1-2 and datedeath_sp1-2 (restricted, cubic spline on calendar time; knots varied for Chinese vs. other cohorts), explag1 (silica exposure from the previous year), cye1_3_1-2 (category indicators for cumulative years of exposure), exp_2_15 (silica exposure accrued between 2 and 15 years prior), cumexplag15 (cumulative silica exposure, 15 year lag), cumyrsexp (cumulative years exposed), xl1cat_5_2-4 (category indicators for exposure in the previous year), sil2_4_2-3 (category indicators for cumulative silica exposure, 2 year lag), datec (calendar time, centered), datesq and datecu (datec squared and cubed), cumexp (cumulative exposure), atwork_2yrs (category indicator of whether individual was employed two years prior)

[b] *In applicable studies only where these factors varied (e.g. race was not used in the models among the Chinese cohorts)*

eTable 2: Hazard ratios and 95% confidence intervals from a Cox proportional hazards model for the association between active employment and the all-cause-mortality.

Lag	HR[a] (95% CI)
0	0.426 (0.404, 0.449)
1	0.562 (0.535, 0.59)
2	0.597 (0.569, 0.627)
3	0.635 (0.606, 0.399)
4	0.659 (0.629, 0.689)

a Hazard ratio for comparing rates of all-cause mortality among actively-employed person-time and person-time not employed in study cohort. Adjusted for log-cumulative exposure (1 year lag, linear term), sex, race, date of birth (linear term), calendar period (linear term), and study (stratification).

eTable 3: Coefficients and 95% confidence intervals from a Weibull accelerated failure time model for the association between log-cumulative exposure and the time-to-leave-employment

Lag	Coefficient[a] (95% CI)
1	-0.0030 (-0.00338, -0.00252)
2	-0.0032 (-0.00360, -0.00274)
5	-0.0028 (-0.00323, -0.00238)
10	-0.0026 (-0.00301, -0.00217)
15	-0.0031 (-0.00354, -0.00273)

a Coefficient for the change in the log-time-to-leave-employment per unit of log-cumulative exposure, assuming a Weibull distribution for the time-to-leave. Adjusted for sex, race, date of birth (linear term), calendar period (restricted cubic spline), and study (study specific Weibull shape parameters and study specific intercepts).

eTable 4: Coefficients and 95% confidence intervals from a Weibull accelerated failure time model for the association between ever- silicosis and the time-to-leave-employment

Lag	Coefficient[a] (95% CI)
0	-0.142 (-0.152, -0.132)
1	-0.073 (-0.079, -0.068)
2	-0.049 (-0.052, -0.045)
3	-0.030 (-0.033, -0.027)
4	-0.021 (-0.023, -0.019)

a Coefficient for the change in the log-time-to-leave-employment for those who have ever been diagnosed with silicosis, compared to those who have not been diagnosed with silicosis, assuming a Weibull distribution for the time-to-leave. Adjusted for log-cumulative exposure (linear term, lagged 1 year longer than employment lag), silicosis prior to baseline, sex, race, date of birth (linear term), calendar period (linear term), and study (study specific Weibull shape parameters and study specific intercepts).

eTable 5: Deaths per 1000 under no intervention and deaths delayed or prevented by age 80 per 1000 workers in China, considering possible time-varying confounding by silicosis.

		Deaths delayed or prevented per 1000 workers	
	Silicosis as confounder	50 ug/m^3	Eliminate exposure
Lung cancer	Yes	0.863 (-2.57, 4.3)	1.28 (-2.28, 4.85)
	No	0.386 (-3.49, 4.26)	0.849 (-2.98, 4.68)
Other causes	Yes	8.06 (-5.62, 21.7)	10.4 (-2.16, 23)
	No	9.92 (-5.15, 25)	12.0 (-3.02, 27.1)
All causes	Yes	8.92 (-5.12, 23)	11.7 (-1.15, 24.6)
	No	10.3 (-4.25, 24.9)	12.9 (-2.21, 28.0)

eTable 6: Cumulative exposure and proportion of person-time exposed in the data and the predicted values under the natural course.

	Mean cumulative exposure (mg/m^3-years)[a]				Time exposed	
	Mean		Standard deviation		Proportion	
	Observed	G-formula	Observed	G-formula	Observed	G-formula
US diatomaceous	2.0	3.2	3.4	6.1	0.28	0.35
South Africa gold	4.2	4.4	1.5	1.5	0.01	0.10
US gold	0.5	0.6	0.8	0.8	0.16	0.19
Australia gold	11.4	11.7	7.6	7.7	0.45	0.48
US granite	1.8	2.3	3.4	3.9	0.50	0.57
Finnish granite	11.4	10.4	17.0	13.6	0.31	0.37
US industrial sand	0.8	1.2	2.0	2.4	0.33	0.38
China Tungsten	14.8	14.8	25.7	23.8	0.65	0.70
China pottery	3.9	3.8	7.0	6.8	0.62	0.68
China tin	3.9	3.2	6.2	5.3	0.73	0.78
Overall	8.1	8.0	18.2	16.9	0.54	0.59

[a] Note that estimates differ from those given for the study data in Table 1 of Steenland et al. [4] because we report the mean and standard deviation across the person-time-at-risk, rather than the median at the end of follow-up.

eTable 7: Crude mortality rates per 10,000 person-years: observed data and g-formula natural course

Industry	Lung cancer		All other causes	
	Observed	G-formula	Observed	G-formula
US diatomaceous	11.7	12.3	101.7	100.5
South Africa gold	24.5	26.5	296.1	282.4
US gold	13.6	13.9	154.7	156.4
Australia gold	30.9	31.7	276.6	272.9
US granite	10.3	10.5	135.3	135.6
Finnish granite	13.6	14.6	132.1	129.6
US industrial sand	9.3	9.5	85.2	86.9
China Tungsten	2.8	2.8	71.5	68.7
China pottery	4.4	4.7	77.6	73.7
China tin	7.2	7.0	48.9	47.4
Overall	7.3	7.6	95.7	94.3

eTable 8: Sensitivity analysis results: cubic spline on exposure

	Deaths per 1000	Deaths delayed or prevented per 100 workers				
	No intervention	100 ug/m^3	50 ug/m^3	25 ug/m^3	10 ug/m^3	Eliminate exposure
Lung cancer	55.1 (53, 57.3)	2.07 (-0.839, 4.98)	3.11 (0.0857, 6.12)	3.61 (0.688, 6.54)	4.3 (0.869, 7.73)	4.68 (1.04, 8.33)
Other causes	647 (642, 653)	8.6 (0.82, 16.4)	10.9 (3.37, 18.5)	12.9 (5.21, 20.6)	14.2 (6.1, 22.3)	14.9 (7.15, 22.6)
All causes	703 (697, 708)	10.7 (3.09, 18.3)	14 (6.2, 21.9)	16.5 (8.44, 24.6)	18.5 (9.78, 27.3)	19.6 (10.6, 28.5)

eTable 9: Deaths per 1000 under no intervention and deaths delayed or prevented by age 80 per 1000 workers under assumption of no healthy worker survivor bias

	Deaths per 1000	Deaths delayed or prevented per 1000 workers (95% CI)				
	No intervention	100 μg/m^3	50 μg/m^3	25 μg/m^3	10 μg/m^3	Eliminate exposure
Lung cancer	55.1 (53.1, 57)	1.69 (-1.13, 4.5)	2.11 (-0.703, 4.92)	2.32 (-0.472, 5.11)	2.54 (-0.329, 5.4)	2.8 (0.0463, 5.55)
Other causes	648 (642, 654)	3.38 (-4.41, 11.2)	5.47 (-2.22, 13.2)	6.53 (-1.02, 14.1)	7.48 (-0.329, 15.3)	8.4 (1.28, 15.5)
All causes	703 (698, 709)	5.07 (-2.63, 12.8)	7.58 (0.140, 15.0)	8.85 (1.54, 16.2)	10.0 (2.19, 17.8)	11.2 (4.08, 18.3)

eTable 10: Deaths per 1000 under no intervention and deaths delayed or prevented by age 80 per 1000 workers under alternative assumption of no healthy worker survivor bias in which exposure is allowed to affect employment

	Deaths per 1000	Deaths delayed or prevented per 100 workers				
	No intervention	100 ug/m^3	50 ug/m^3	25 ug/m^3	10 ug/m^3	Eliminate exposure
Lung cancer	55.1 (53.1, 57.1)	1.77 (-1.04, 4.58)	2.32 (-0.51, 5.16)	2.36 (-0.292, 5.01)	2.53 (-0.238, 5.3)	2.71 (-0.0486, 5.47)
Other causes	648 (643, 653)	3.03 (-4.3, 10.4)	5.01 (-2.37, 12.4)	6.13 (-1.63, 13.9)	7.44 (0.169, 14.7)	8.11 (0.321, 15.9)
All causes	703 (698, 708)	4.8 (-2.4, 12)	7.33 (-0.02, 14.7)	8.49 (0.671, 16.3)	9.97 (2.39, 17.5)	10.8 (3.27, 18.4)

References

1. Keil AP, Edwards JK, Richardson DB, Naimi AI, Cole SR. The parametric g-formula for time-to-event data: intuition and a worked example. *Epidemiology* 2014;**25**(6):889-97.
2. Keil AP, Richardson DB. Reassessing the Link between Airborne Arsenic Exposure among Anaconda Copper Smelter Workers and Multiple Causes of Death Using the Parametric g-Formula. *Environ Health Perspect* 2016;**125**(4):608-614.
3. Taubman SL, Robins JM, Mittleman MA, Hernan MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int J Epidemiol* 2009;**38**(6):1599-611.
4. Steenland K, Mannetje A, Boffetta P, Stayner L, Attfield M, Chen J, Dosemeci M, DeKlerk N, Hnizdo E, Koskela R, Checkoway H, IARC. Pooled exposure-response analyses and risk assessment for lung cancer in 10 cohorts of silica-exposed workers: an IARC multicentre study. *Cancer Causes Control* 2001;**12**(9):773-84.