Estimating the Effect of Cumulative Occupational Asbestos Exposure on Time to Lung Cancer Mortality

Using Structural Nested Failure-Time Models to Account for Healthy-Worker Survivor Bias

Ashley I. Naimi,[a] *Stephen R. Cole,*[a] *Michael G. Hudgens,*[b] *and David B. Richardson*[a]

Background: Previous estimates of the effect of occupational asbestos on lung cancer mortality have been obtained by using methods that are subject to the healthy-worker survivor bias. G-estimation of a structural nested model provides consistent exposure effect estimates under this bias.

Methods: We estimated the effect of cumulative asbestos exposure on lung cancer mortality in a cohort comprising 2564 textile factory workers who were followed from January 1940 to December 2001.

Results: At entry, median age was 23 years, with 42% of the cohort being women and 20% nonwhite. During the follow-up period, 15% of person-years were classified as occurring while employed and 13% as occupationally exposed to asbestos. For a 100 fiber-year/ml increase in cumulative asbestos, a Weibull model adjusting for sex, race, birth year, baseline exposure, and age at study entry yielded a survival time ratio of 0.88 (95% confidence interval = 0.83 to 0.93). Further adjustment for work status yielded no practical change. The corresponding survival time ratio obtained using g-estimation of a structural nested model was 0.57 (0.33 to 0.96).

Conclusions: Accounting for the healthy-worker survivor bias resulted in a 35% stronger effect estimate. However, this estimate was considerably less precise. When healthy-worker survivor bias is suspected, methods that account for it should be used.

(*Epidemiology* 2014;25: 246–254)

Submitted 26 April 2013; accepted November 12, 2013.

From the [a]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC; and [b]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC.

The authors report no conflicts of interest.

Supported in part through NIH-NCI grant R01CA117841. Supported by a Doctoral Research Award from the Fonds de Recherche en Santé du Québec (to A.I. N.).

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com). This content is not peer-reviewed or copy-edited; it is the sole responsibility of the author.

Correspondence: Ashley I. Naimi, Department of Epidemiology, CB7415, University of North Carolina, Chapel Hill, NC 27599. E-mail: ashley.naimi@unc.edu.

ISSN: 1044-3983/14/2502-0246
DOI: 10.1097/EDE.0000000000000045

The goal of many epidemiologic studies is to estimate the effect of a time-varying exposure on time to an outcome of interest. In such studies, exposure effect estimates are often prone to time-varying confounding. A time-varying confounder is a variable (1) whose values change over time, (2) that influences exposure at subsequent time points, and (3) that is associated with the time to the outcome of interest. When a time-varying confounder is also related to previous exposure values, standard methods such as Cox regression will not consistently estimate the total (ie, direct and indirect) effect of exposure on the outcome of interest.[1–4] In particular, although standard methods may resolve confounding bias, the net bias may be increased upon adjustment because of collider stratification.[5–7]

Moreover, additional problems arise when there are no exposed or no unexposed persons within a given stratum of a time-varying confounder. This results in a violation of the positivity assumption, which formally requires an exposure probability bounded away from 0 or 1 in all confounder strata and at all time points.[8–10] A canonical example of such time-varying confounding with a positivity violation is the healthy-worker survivor bias, which is depicted in Figure 1.[11–13] This bias arises in situations where employment status at time t, denoted W_t, is associated with the outcome of interest, T. In such a scenario, the effect of an occupationally based exposure X_t may be confounded by employment status, whereas employment status may be affected by previous exposure, denoted X_{t-1}. In addition, persons who leave employment often have no chance of incurring work-based exposure at subsequent time points, resulting in a violation of the positivity assumption. Because persons who leave active employment are no longer exposed, and employment status may be associated with exposure at previous time points, consistent exposure effect estimates cannot be obtained using standard analytic techniques.[3,14–16]

Although standard methods provide inconsistent estimates under these conditions, g-estimation of a structural nested accelerated failure-time model can be used to appropriately adjust for the healthy-worker survivor bias.[1,3,17,18] To date, only one study[19] has used structural nested models to account for the healthy-worker survivor bias, in which an occupational

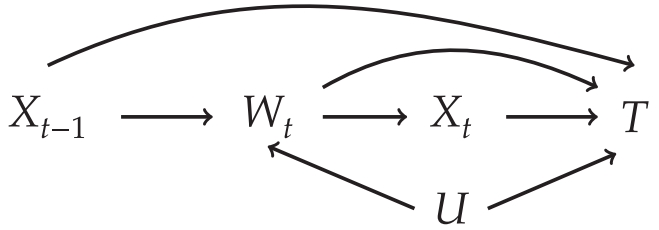

FIGURE 1. Causal diagram representing the healthy-worker survivor bias. We let *t* index years on study, *X* represent continuous asbestos exposure, *W* represent employment status, *U* represent a common cause of *W* and *T*, and *T* represent survival time.

exposure metric originally measured on a continuous scale was dichotomized. Another study[20] has implemented structural nested models to estimate the effect of erythropoietin dose (on a continuous scale) on mortality in patients with end-stage renal disease. No studies have used structural nested models to account for healthy-worker survivor bias for the relation between a continuous occupational exposure and mortality. The purpose of this article is to (1) describe the application of structural nested accelerated failure-time models with a continuous exposure variable, using data from a cohort comprising 2564 asbestos textile factory workers followed for lung cancer mortality during 100,765 person-years between 1940 and 2001 and (2) compare effect estimates obtained using this method to those obtained using more standard parametric accelerated failure-time models.[21]

METHODS

Study Cohort

The South Carolina Chrysotile Asbestos study was an occupational cohort study of the relation between workplace asbestos exposure and lung cancer mortality during a 60-year period. We previously have found evidence that the relations depicted in Figure 1 are present in this cohort.[22] The cohort consisted of 3072 persons who worked in the plant for at least 1 month between 1 January 1940 and 31 December 1965.[23] Of these, 508 persons had been working in the plant before 1 January 1940 and were excluded from the study, leaving an incident-hire cohort of 2564 workers followed for 100,765 person-years for vital status and cause of death. Follow-up ended at death, loss to follow-up, or administrative censoring on 31 December 2001. Date of birth, sex, and race (white vs. nonwhite) were ascertained from company personnel records. This study was conducted on deidentified existing records and therefore not considered human subjects research.

As with previous research in this cohort,[23] the primary outcome is lung cancer mortality. Mortality data were obtained from the National Death Index, with lung cancer death defined by the appropriate International Classification of Disease codes, as described in detail elsewhere.[22]

Levels of asbestos exposure were assigned using a job-exposure matrix. Annual exposure levels, in fiber-years/ml, were calculated as the product of duration of exposure in a given calendar year and the department-, task-, and calendar-period specific average concentration of chrysotile fibers longer than 5 μm/ml air to which a worker was exposed in that year. This exposure metric was chosen on the basis of previous research[23] and is described in detail elsewhere.[22]

Statistical Methods

The information from the cohort provided data for each of the 2564 study participants. Let $Y = \min(T, C)$ denote the years from study entry to lung cancer mortality (denoted T, with event indicator $\delta = 1$) or censoring because of deaths from other causes, drop out, or study completion (denoted C, $\delta = 0$). We assume noninformative censoring conditional on measured covariates and illustrate how to account for it in Appendix 1 (for standard methods) and Appendix 2 (for structural nested models). Let $\overline{X_t}$ denote the set of annual exposure values, in units of 100 fiber-years/ml, representing a person's exposure status during the course of follow-up, up to year t. Similarly, let $\overline{W_t}$ denote the set of indicators of whether the worker was actively employed at the facility under study in each year during the course of follow-up, up to time t. For example, for a worker who left employment mid-year in their third year on study during 5 years of follow-up, $\overline{W_4} \equiv \{W_0,\ldots,W_4\} = \{1, 1, 1, 0, 0\}$. Finally, L is a vector of time-fixed covariates consisting of age at study entry (age), birth year (byr), and binary indicators for sex (female = 1) and race (nonwhite = 1).

We assess the relation between occupational asbestos exposure and lung cancer mortality with two accelerated failure-time modeling strategies: a standard model and a structural nested model. Both models estimate the ratio of median survival times (relative times) for a 100 fiber-year/ml increase in cumulative occupational asbestos exposure, lagged by 10 years. We lagged the exposure by 10 years to account for an empirical induction period[24] based on previous literature.[23,25] Specifically, we defined a cumulative exposure metric with a 10-year lag as $A_t = \sum_{j=0}^{t-9} X_j$ if $t \geq 9$, and $A_t = 0$ otherwise. For all models, we used time from entry into the study as the time scale. In all models, age at study entry, birth year, and pre-entry exposure were entered as restricted quadratic splines with knots at the 5th, 28th, 50th, 73rd, and 95th percentiles of the variable's distribution.[26] For each method, relative times were chosen as measures of the effect of asbestos exposure, 95% confidence intervals as measures of precision, and confidence limit ratios (defined as the ratio of the upper and lower confidence bounds)[27] as measures of efficiency. Comparing efficiency measures allows for an assessment of how well each method uses the information in the sample to estimate relative times. We opt for relative times rather than hazard ratios because of the simple interpretation of the former[28(p 450)] and the inherent selection bias associated with the latter.[29]

The standard parametric accelerated failure time-model[30(p 321)] took the form:

$$F^{-1}\{P(T \le t \mid Z_t = z_t)\} = \int_0^t \exp(\beta' z_u)\mathrm{d}u \qquad \text{Model 1}$$

where $F^{-1}\{\cdot\}$ is the inverse of a Weibull cumulative distribution function and Z_t is a vector of covariates containing the exposure and measured confounders. In this model, the exponentiated β coefficient for A_t can be interpreted as the associational survival time ratio for a 100 fiber-year/ml increase in A_t, adjusting for all other covariates in the model. For example, $\exp(\beta)$ can be interpreted as the ratio of the (eg, median) survival time for individuals exposed to α + 100 fiber-years/ml of cumulative asbestos relative to α fiber-years/ml of cumulative asbestos. We estimated parameters for model 1 with maximum likelihood accounting for right censoring and time-varying covariates.[31] Technical details are provided in Appendix 1. Even with a model adjusted for all relevant confounders, the maximum likelihood estimator will be inconsistent for the exposure effect when the conditions leading to the healthy-worker survivor bias are present.[1–3]

We also specified a structural nested accelerated failure-time model as:

$$T^{\bar{0}} = \int_0^T \exp(\psi A_u)\mathrm{d}u \qquad \text{Model 2}$$

where $T^{\bar{0}}$ is the potential survival time that would have been observed under no exposure. In this model, $\exp(-\psi)$ is the causal survival time ratio for a 100 fiber-year/ml increase in A_t. The parameter ψ for model 2 was estimated using g-estimation with a line search method (Appendix 2). We restricted our line search to the parameter space defined by $\{0,\ldots,2\}$ by 0.05, and denote this set of values $\widetilde{\psi}$. Excluding negative values of the parameter space assumes that asbestos is not protective for lung cancer mortality, which is well-justified by previous research.[32]

In contrast to standard methods, which require specification of the confounder–outcome relation, g-estimation adjusts for confounding by modeling the confounder–exposure relation.[33] This allowed for more flexible confounder adjustment in these data. To model the exposure, we categorized X_t into 11 categories (one category for $X_t = 0$, and deciles for $X_t > 0$), and used an ordinal logistic regression model adjusting for race, sex, age at study entry, and birth year. Inverse probability weights were used to account for right censoring because of drop out or competing risks.[1–3] Technical details are provided in Appendix 2. Assuming a correctly specified structural nested model (model 2) and no unmeasured confounding (Equation A2.1), this g-estimator will be consistent for the exposure effect, whether or not the conditions that lead to healthy-worker survivor bias are present.[1–3] We used SAS release 9.3 for all analyses (SAS Institute, Cary, NC). A version of the code to fit this structural nested accelerated failure-time model via g-estimation is provided elsewhere.[34]

RESULTS

Table 1 presents study characteristics of 2564 workers. During the course of 100,765 person-years of follow-up, 15% of person-years were spent at work and 13% were classified as exposed to any asbestos. Furthermore, 89% of workers (n = 2285) left active employment alive, with the remaining 11% either lost to follow-up (n = 236), or having died of lung cancer (n = 4) or a competing cause of death (n = 39) while actively employed. An additional 14 workers were lost to follow-up after leaving active employment, for a total of 250 (9.7% of 2564) lost to follow-up. A total of 142 workers (5.5%) were classified as having died because of lung cancer.

For a 100 fiber-year/ml increase in cumulative asbestos, the standard accelerated failure-time model adjusting for sex, race, birth year, baseline exposure, and age at study entry yielded an estimated survival time ratio of 0.88 (95% confidence interval = 0.83 to 0.93). Further adjustment for work status yielded an estimated survival time ratio of 0.89 (0.82 to 0.96). These models yielded confidence limit ratios of 1.12 and 1.17, respectively (Table 2).

Figure 2 displays the plot of the *Z*-statistics for each value of $\widetilde{\psi}$ (as described in Appendix 2), with the gray lines representing results from 200 bootstrap resamples. The value of $\widetilde{\psi}$ that minimizes the *Z*-statistic from a test of the independence between the exposure and the potential outcomes is the estimate $\widehat{\psi}$. As shown in Figure 2, the *Z*-statistic crosses zero at $\widetilde{\psi} = 0.57$, yielding an estimated survival time ratio of $\exp\{-\widehat{\psi}\} = 0.57$ per 100 fiber-year/ml increase in cumulative exposure to asbestos, with 95% bootstrap confidence intervals

TABLE 1. Characteristics[a] of 2564 Persons in the South Carolina Chrysotile Asbestos Cohort at Enrollment and During 100,765 Person-Years of Follow-Up Between 1940 and 2001

Characteristics	First Year of Follow-up (n = 2,564)	Entire Course of Follow-up (No. Person-Years = 100,765)
Age	23 (20–31)	22 (19–28)
Birth year	1922 (1915–1926)	1922 (1916–1927)
Women; n (%)	1,077 (42)	42,416 (42)
Nonwhite race; n (%)	512 (20)	18,305 (18)
Working; n (%)	2,564 (100)	15,537 (15)
Asbestos exposure (fiber-years/ml)		
Any; n (%)	2,314 (90)	12,781 (13)
Continuous (X_t)[b]	1.7 (0.9–3.2)	3.4 (1.6–5.0)
Cumulative (A_t)	NA[c]	1.7 (0–8.2)

[a]Median (quartiles) unless otherwise stated.
[b]Among person-years with any exposure.
[c]Ten-year lagged cumulative exposure = 0 during the first year of follow-up for all persons in the study.
NA indicates not available.

TABLE 2. Survival Time Ratios (95% CIs) for the Association Between 10-Year Lagged Cumulative Asbestos Exposure for Standard and Structural Nested Model Analyses

Accelerated Failure-Time Model Model	Time Ratio (95% CI)	CLR[a]
Work status unadjusted	0.88 (0.83–0.93)	1.12
Work status adjusted	0.89 (0.82–0.96)	1.17
Structural nested	0.57 (0.33–0.96)[b]	2.91
	(0.32–1.00)[c]	3.13
	(0.17–0.95)[d]	5.59

[a]CLR (upper limit divided by lower limit).
[b]Bootstrap Wald CIs where $SE(\widehat{\psi}) = SD(\psi_k)$ with $K = 200$ resamples.
[c]Test-based CIs for $\exp\{-\psi\}$ chosen such that $Z = \pm 1.96$.
[d]Slope-based Wald CIs with $SE(\widehat{\psi}) = \left(d^{-2}\right)^{1/2}$, where d is obtained from the slope of a local linear regression of Z on $\widetilde{\psi}$.
CI indicates confidence interval; CLR, confidence limit ratio.

of (0.33 to 0.96) and a confidence limit ratio of 2.91. As shown in Table 2, bootstrap- and test-based 95% confidence intervals (Appendix 2) were similar. The survival time ratio obtained from the structural nested model is 35% stronger than the ratio obtained from both Weibull accelerated failure-time models (Table 2). Compared with the fully adjusted parametric model, the structural nested model was 60% less efficient as determined by 95% bootstrap confidence intervals. See the eAppendix (http://links.lww.com/EDE/A751) for further discussion on the methods used to obtain confidence intervals.

DISCUSSION

Using structural nested accelerated failure-time models, we estimated that an increase in 100 fiber-years/ml of work-based cumulative asbestos is associated with a 43% shorter time to lung cancer mortality. This was in contrast to an estimated 12% and 11% shorter time to lung cancer mortality obtained from the work-status–unadjusted and work-status–adjusted parametric accelerated failure-time models, respectively.

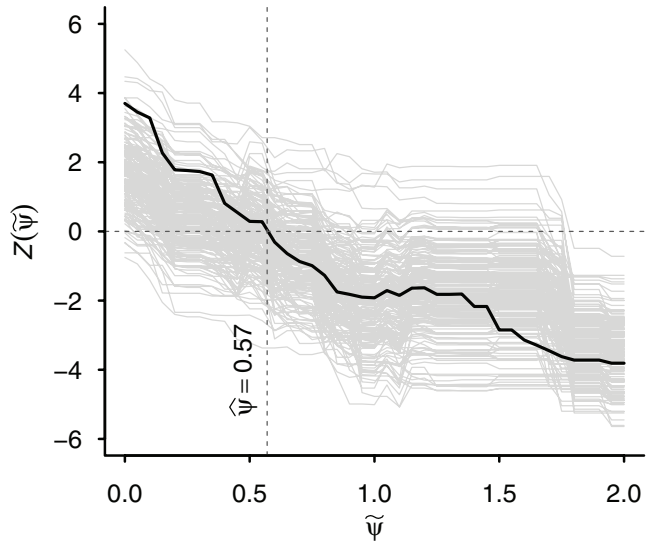

FIGURE 2. Plot of the $Z(\widetilde{\psi})$ test statistic by $\widetilde{\psi}$ including a horizontal reference line at zero and vertical reference line at $\widehat{\psi}$. Gray lines represent results for each of the 200 bootstrap resamples.

There are several possible explanations for the difference in point estimates between standard and structural nested models. First, the difference may be attributable to healthy-worker survivor bias. Second, structural nested models are semi-parametric: nonparametric with respect to both the distribution of the baseline survival time,[35] and the functional form of the relation between measured confounders (eg, age and birth year) and lung cancer survival time,[33] and parametric (model 2) with respect to the functional form between cumulative asbestos exposure and lung cancer survival time. Standard parametric accelerated failure-time models, however, assume a priori that the baseline survival times follow a specified distribution (eg, Weibull), and assumes a parametric form for the relation between all covariates (exposure and confounders) and the outcome.

In theory, part of the observed difference may be because of residual confounding caused by mis-specification of measured confounders.[36(p 198)] With structural nested models, one typically adjusts for confounding by modeling the exposure as a function of measured confounders (Appendix 2). With a common exposure, one can leverage the available information to specify a flexible model for confounding (by including, eg, splines, fractional polynomials, and first- and higher-order interactions). In contrast, standard regression models typically adjust for confounding by modeling the outcome as a function of exposure and measured confounders. With a less common outcome, model stability can become problematic with flexible confounder specification.

Indeed, in the asbestos cohort study data, we were able to obtain convergence while adjusting the structural nested model (using g-estimation) for all baseline confounders with all first- and second-order interactions. By using the standard parametric accelerated failure-time models, we could adjust only for all baseline confounders and select first-order interactions (ie, the maximization algorithm failed to converge to a unique solution when all first- and second-order interactions were included). Nevertheless, removing from the structural nested model interaction terms that were not in the parametric model yielded negligible changes in the estimated survival time ratio, thus ruling out residual confounding as a possible explanation of the differences between our two modeling strategies.

Given the distinctions between standard and structural nested accelerated failure-time models, it can be argued that the parameter estimates from structural nested models would be preferable: (1) better confounder control is always desired when the goal is unbiased point estimation; (2) semi-parametric methods make less restrictive assumptions about the supposed

mechanisms that generated the data; and (3) structural nested models can account for time-varying confounding when the positivity assumption is violated (as with the healthy-worker survivor bias).

However, confidence intervals for the parameters from the structural model were much larger compared with the confidence intervals from the parametric models, as can be seen in Table 2. This is the expected result of the fact that: (1) semi-parametric methods are known to be less efficient than their correctly specified parametric counterparts[37]; (2) the chosen functional form of the potential outcomes in our exposure model (Appendix 2) is known to result in an efficiency loss[2]; and (3) artificial censoring, required for consistent estimation of structural nested model parameters when administrative censoring is present (Appendix 2), reduces the effective number of events.[20] Indeed, as Joffe et al[20] have shown, this latter issue can curtail the estimation of causal effects using structural nested failure-time models, particularly when the exposure is continuous. We were able to circumvent many of these issues by restricting our grid search to a non-negative portion of the parameter space. This restriction corresponds to the well-justified assumption that the average causal effect of asbestos on lung cancer mortality is not protective.[32] In other settings where the direction of effect is less clear, use of structural nested failure-time models may be more complicated.

Any statistical method must rely on several identifiability assumptions to make valid inferences. In particular, we assume no unmeasured confounding.[2,38] In our case, we did not have information on individual smoking status, socioeconomic position, or intermittent time off work, which is a limitation.

In addition, we assume counterfactual consistency, which states that for a given person, the potential outcome that would have been observed under the observed exposure is, in fact, the observed outcome.[39–43] This assumption is better satisfied with an exposure contrast that can be construed as an intervention.[39–41] In our study, we defined the exposure as the cumulative number of chrysotile asbestos fibers longer than 5 µm/ml of air to which a worker was exposed in an asbestos textile factory, in units of 100 fiber-years/ml. Our exposure contrast compares the survival time under some arbitrary level of cumulative exposure over time $\overline{a}_t$ with the survival time that would have been observed under a cumulative exposure level $\overline{a}_t + 100$ (ie, a 100 fiber-years/ml increase). Such a change in the exposure value could ostensibly be achieved through a number of interventions. In the South Carolina factory, increasing the amount of workplace ventilation and exhaust was the intervention of choice to reduce exposure.[44] Alternatively, use of a self-contained breathing apparatus (as done in the asbestos abatement industry) could reduce exposure corresponding to the exposure contrast used in this study.[45] Counterfactual consistency requires that interventions producing an identical decrease in the exposure would have the same causal effect. As with all analyses using observational data to estimate causal effects, this assumption cannot be verified and should be made with caution.

We further assume no interference and correct model specification. The former assumption implies that no worker's potential outcome is influenced by another worker's exposure history,[46] which is likely in our setting. For standard accelerated failure-time models, the latter assumption requires a correctly chosen: (1) distribution for baseline survival time; (2) link function relating the outcome to the model covariates (in our case, *exp*(), integrated over *t*); and (3) the functional form of all covariates involved. For the structural nested model, the latter assumption requires only a correctly specified link function (*exp*(), integrated over *t*)), and a correct functional form of the relation between cumulative asbestos exposure and time to lung cancer mortality (or a meaningful projection thereof).[47]

Finally, the positivity (or experimental-treatment-assignment) assumption is a central requirement for valid causal inferences using observational data. Positivity requires exposed and unexposed observations in all confounder strata at all time points.[8–10] The healthy-worker survivor bias results in a systematic (as opposed to random) violation of positivity[16]: persons who have terminated active employment cannot be exposed to work-based asbestos.

Systematic positivity violations can be problematic for two reasons. First, it is possible that the data contain information that can be used to make valid causal contrasts, but because of a lack of positivity, some methods fail (whereas other methods succeed) to estimate such contrasts. For example, inverse probability weighting estimators of the parameters of a marginal structural model are not consistent when the positivity assumption is violated.[16,48] Other marginal structural model estimators (g-computation,[49] augmented inverse probability weighting,[50] or targeted maximum likelihood estimators),[51] or other methods (parametric g-formula[14] or g-estimation of a structural nested model)[17] can provide consistent (ie, unbiased) estimates, even under positivity violations. However, positivity violations may be such that there is no relevant information upon which one can make a valid causal contrast.[10,52–54] In such a scenario, researchers must avoid methods that provide point estimates for a causal contrast that is unsupported by data, of little meaning or public health relevance, or both. In our specific case, we are estimating the effect of a 100 fiber-year/ml change in work-based cumulative asbestos exposure on lung cancer survival. Because of our use of g-estimation (Appendix 2), this contrast is supported by the data.

We assessed the relation between cumulative asbestos exposure and lung cancer mortality using standard and structural nested accelerated failure-time models. The latter can account for the healthy-worker survivor bias. Despite limitations, our analysis is strengthened by the use of a large cohort with well-characterized mortality and less than 10% loss to follow-up during a 60-year period. In our example, the

structural accelerated failure-time model yielded a survival time ratio that was 35% stronger than the estimates from the standard accelerated failure-time models. This is suggestive of bias in previous analyses with these data, and the magnitude is such that it underscores importance of accounting for healthy-worker survivor bias in this setting. When the healthy-worker survivor bias is suspected, researchers should opt for appropriate methods (such as structural nested models) to obtain unbiased estimates of the effect of a work-based exposure on a health outcome.

ACKNOWLEDGMENTS

We thank Steven B. Wing and M. Alan Brookhart for comments on earlier versions of the article. We thank Marshall M. Joffe and Erica E.M. Moodie for expert advice.

REFERENCES

1. Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, eds. *Health Service Research Methodology: A Focus on AIDS*. Washington, D.C.: U.S. Public Health Service, National Center for Health Services Research; 1989:113–159.
2. Robins JM, Blevins D, Ritter G, Wulfsohn M. G-estimation of the effect of prophylaxis therapy for *Pneumocystis carinii* pneumonia on the survival of AIDS patients (erratum in Epidemiology 1993; 3:189). *Epidemiology*. 1992;3:319–336.
3. Robins JM. Analytic methods for estimating HIV-treatment and cofactor effects. In: Ostrow D, Kessle R, eds. *Methodological Issues of AIDS Behavioral Research*. New York, N.Y.: Plenum; 1993:213–287.
4. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550–560.
5. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–625.
6. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003;14:300–306.
7. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol*. 2010;39:417–420.
8. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006;60:578–586.
9. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168:656–664.
10. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol*. 2010;171:674–677; discussion 678.
11. Eisen E, Picciotto S, Robins J. Healthy worker effect. In: El-Shaarawi A, Piegorsch W, eds. *Encyclopedia of Environmetrics*. New York, N.Y.: John-Wiley & Sons; 2013.
12. Arrighi HM, Hertz-Picciotto I. Definitions, sources, magnitude, effect modifiers, and strategies of reduction of the healthy worker effect. *J Occup Med*. 1993;35:890–892.
13. Arrighi HM, Hertz-Picciotto I. The evolving concept of the healthy worker survivor effect. *Epidemiology*. 1994;5:189–196.
14. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathemat Model*. 1986;7:1393–1512.
15. Robins JM. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chronic Dis*. 1987;40(suppl 2):139S–161S.
16. Naimi AI, Cole SR, Westreich D, Richardson D. A comparison of methods to estimate the hazard ratio under conditions of time-varying confounding with nonpositivity. *Epidemiology*. 2011;22:718–723.
17. Robins JM. Structural nested failure time models. In: Andersen P, Keiding N, eds. *The Encyclopedia of Biostatistics*. Chichester, United Kingdom: John Wiley and Sons; 1998.
18. Hernán MA, Cole SR, Margolick J, Cohen M, Robins JM. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiol Drug Saf*. 2005;14:477–491.
19. Chevrier J, Picciotto S, Eisen EA. A comparison of standard methods with g-estimation of accelerated failure-time models to address the healthy-worker survivor effect: application in a cohort of autoworkers exposed to metalworking fluids. *Epidemiology*. 2012;23:212–219.
20. Joffe MM, Yang WP, Feldman H. G-estimation and artificial censoring: problems, challenges, and applications. *Biometrics*. 2012;68:275–286.
21. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis—an introduction to concepts and methods. *Br J Cancer*. 2003;89:431–436.
22. Naimi AI, Cole SR, Hudgens MG, Brookhart MA, Richardson DB. Assessing the component associations of the healthy worker survivor bias: occupational asbestos exposure and lung cancer mortality. *Ann Epidemiol.* 2013;23:334–341.
23. Hein MJ, Stayner LT, Lehman E, Dement JM. Follow-up study of chrysotile textile workers: cohort mortality and exposure-response. *Occup Environ Med*. 2007;64:616–625.
24. Rothman KJ. Induction and latent periods. *Am J Epidemiol*. 1981;114:253–259.
25. Richardson DB, Cole SR, Chu H, Langholz B. Lagging exposure information in cumulative exposure-response analyses. *Am J Epidemiol*. 2011;174:1416–1422.
26. Howe CJ, Cole SR, Westreich DJ, Greenland S, Napravnik S, Eron JJ Jr. Splines for trend analysis and continuous confounder control. *Epidemiology*. 2011;22:874–875.
27. Poole C. Low P-values or narrow confidence intervals: which are more durable? *Epidemiology*. 2001;12:291–294.
28. Reid N. A conversation with Sir David Cox. *Stat Sci*. 1994;9:439–455.
29. Hernán MA. The hazards of hazard ratios. *Epidemiology*. 2010;21:13–15.
30. Lawless JF. *Statistical Models and Methods for Lifetime Data. Wiley Series in Probability and Statistics*. Hoboken, N.J.: Wiley-Interscience; 2003.
31. Cox C, Chu H, Schneider MF, Muñoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Stat Med*. 2007;26:4352–4374.
32. Tweedale G. Asbestos and its lethal legacy. *Nat Rev Cancer*. 2002;2:311–315.
33. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*. 1992;48:479–495.
34. Naimi AI, Richardson DB, Cole SR. Causal inference in occupational epidemiology: accounting for the healthy worker effect using structural nested models. *Am J Epidemiol*. 2013;178:1681–1686.
35. Robins JM, Tsiatis AA. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika*. 1992;79:311–319.
36. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. New York, N.Y.: Wolters Kluwer–Lippincott Williams & Wilkins; 2008.
37. Newey WK. Semiparametric efficiency bounds. *J Appl Econ*. 1990;5:99–135.
38. Robins JM. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*. 1992;79:321–334.
39. Hernán MA. Invited commentary: hypothetical interventions to define causal effects—afterthought or prerequisite? *Am J Epidemiol*. 2005;162:618–620; discussion 621.
40. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *Int J Obes (Lond)*. 2008;32(suppl 3):S8–S14.
41. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009;20:3–5.
42. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology*. 2009;20:880–883.
43. Pearl J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology*. 2010;21:872–875.
44. Dement J. *Estimation of Dose and Evaluation of Dose-Response in a Retrospective Cohort Mortality Study of Chrysotile Asbestos Textile Workers*. Chapel Hill, N.C.: University of North Carolina; 1980.
45. Noonan G, Linn H, Reed L. *A Guide to Respiratory Protection for the Asbestos Abatement Industry*. Morgantown, W.Va.: U.S. Environmental Protection Agency; 1986.
46. Hudgens MG, Halloran ME. Toward causal inference with interference. *J Am Stat Assoc*. 2008;103:832–842.

47. Hubbard AE, Ahern J, Fleischer NL, et al. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*. 2010;21:467–474.
48. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*. 2012;21:31–54.
49. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol*. 2011;173:731–738.
50. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89:846–866.
51. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2.
52. Messer LC, Oakes JM, Mason S. Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding. *Am J Epidemiol*. 2010;171:664–673.
53. Oakes JM. Commentary: advancing neighbourhood-effects research—selection, inferential support, and structural confounding. *Int J Epidemiol*. 2006;35:643–647.
54. Moore KL, Neugebauer R, van der Laan MJ, Tager IB. Causal inference in epidemiological studies with strong confounding. *Stat Med*. 2012;31:1380–1404.
55. Allison PD. *Survival Analysis using SAS: A Practical Guide*. Cary, N.C.: SAS Institute; 2010.
56. Hurley PJ. *A Concise Introduction to Logic*. 11th ed. Belmont, Calif.: Wadsworth Pub. Co.; 2011.
57. Mark SD, Robins JM. Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Stat Med*. 1993;12:1605–1628.
58. Efron B. *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability; 57*. Boca Raton, Fla.: Chapman & Hall/CRC; 1993.
59. Witteman JC, D'Agostino RB, Stijnen T, et al. G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Heart Study. *Am J Epidemiol*. 1998;148:390–401.
60. McCullagh P. Regression models for ordinal data. *J R Stat Soc Series B Stat Methodol*. 1980;42:109–142.
61. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13–22.
62. Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Commun Stat Simul Comput*. 1994;23:939–951.
63. Joffe MM. Administrative and artificial censoring in censored regression models. *Stat Med*. 2001;20:2287–2304.
64. Keiding N, Filiberti M, Esbjerg S, Robins JM, Jacobsen N. The graft versus leukemia effect after bone marrow transplantation: a case study using structural nested failure time models. *Biometrics*. 1999;55:23–28.
65. Cain LE, Cole SR. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. *Stat Med*. 2009;28:1725–1738.

APPENDIX 1: TECHNICAL DETAILS FOR THE TIME-VARYING WEIBULL ACCELERATED FAILURE TIME MODEL

We let $f_T(t)$ and $S_T(t)$ denote the probability density and survival functions for T~Weibull (φ,σ), defined as:

$$f_T(t) = \frac{1}{\sigma t}\exp\{-\varphi_t\}t^{\frac{1}{\sigma}}\exp\{\exp(-\varphi_t)t^{\frac{1}{\sigma}}\}$$

and

$$S_T(t) = 1 - \Gamma\{\exp(-\varphi_t)t^{\frac{1}{\sigma}}\}$$

respectively, where $\varphi_t = \boldsymbol{\beta}' z_t$, and where $\Gamma\{\cdot\}$ is the cumulative distribution function for a generalized gamma distribution defined by Cox et al.[31] For observation time $Y = \min(T,C)$, we assume that conditional on measured confounders, the event time T is independent of censoring time C. For the Weibull accelerated failure-time model accounting for right censoring and time-varying covariates, each person's contribution to the log-likelihood is

$$L_i(\varphi_y,\sigma) = \left\{\prod_{y=0}^{\lfloor Y_i\rfloor}\frac{S_T(y;\varphi_y)}{S_T\left[(y-1)\wedge 0;\varphi_y\right]}\right\}\left\{\frac{f_T(Y_i;\varphi_y)}{f_T(\lfloor Y_i\rfloor;\varphi_y)}\right\}^{\delta_i}\left\{\frac{S_T(Y_i;\varphi_y)}{S_T(\lfloor Y_i\rfloor;\varphi_y)}\right\}^{(1-\delta_i)}$$

where $\lfloor\cdot\rfloor$ denotes the floor function (eg, $\lfloor 5.7\rfloor = 5$), and $a \wedge b = \min(a,b)$. In this equation, the product is taken over integer values of y because annual measures were taken. Thus, all persons contribute to the first term in the likelihood, lung cancer deaths ($\delta_i = 1$) contribute to the second term, and censored observations ($\delta_i = 0$) contribute to the third term.[55(p 220)] We fit this model using a SAS NLMIXED program provided by Cox et al[31] to estimate parametric survival models with time-varying covariates.

APPENDIX 2: G-ESTIMATION

The g-estimation algorithm is premised on the assumption of no unmeasured confounding.[2,38] This assumption states that, at time t, and conditional on past confounders and exposure history, the exposure received is independent of the potential outcomes. No unmeasured confounding implies:

$$E\left\{X_t \mid \overline{X}_{(t-1)}, \boldsymbol{V_t}, T^{\bar{a}}\right\} = E\left\{X_t \mid \overline{X}_{(t-1)}, \boldsymbol{V_t}\right\}, T > t \quad \text{(A2.1)}$$

where E denotes the expectation operator, $\overline{X}_{(t-1)}$ is the exposure history up to and including time $t-1$, and $\boldsymbol{V}_t$ is a vector of time-fixed and time-varying covariates that renders the assumption of no unmeasured confounding as close to true as possible (in our case, the time-fixed variables in $\boldsymbol{L}$ and time-varying W_t). This assumption implies that the exposure allocation mechanism can be thought of as a sequentially randomized (or nested) process: roughly speaking, within strata defined by the joint distribution of $\left[\overline{X}_{(t-1)},\boldsymbol{V_t}\right]$ over all time points, exposure allocation occurs at random (or, by implication, is independent of the potential outcomes).

The g-estimation algorithm is then implemented as follows:

1. Specify a model for Equation A2.1.
2. Specify a set of parameter values denoted $\widetilde{\psi}$ that are likely to span the effect of interest with an appropriate increment

to carry out a line search (eg, 0 to 2 by 0.05) that will be used to estimate ψ.

3. For each element in the set $\widetilde{\psi}$, use model 2 in the main text to impute $T^{\bar{0}}$, which we rewrite as $T(\widetilde{\psi})$ to make the dependence on ψ clear. Thus, $T(\widetilde{\psi})$ denotes the potential survival time under no exposure for each element in $\widetilde{\psi}$.
4. For each element in the set $\widetilde{\psi}$, use the imputed survival times $T(\widetilde{\psi})$ to test whether exposure at time t is conditionally independent of the potential outcomes given $\overline{X}_{(t-1)}, \mathbf{V}_t$. We can test this independence using a Wald, score, or likelihood ratio test. A set of 41 $\widetilde{\psi}$ values (0 to 2 by 0.05) returns a set of 41 different test statistics, denoted $Z(\widetilde{\psi})$.
5. We can obtain our parameter estimate as follows: Assuming no unmeasured confounding holds, the value of $\widetilde{\psi}$ for which the test statistic $Z(\widetilde{\psi}) = 0$ is the value at which we have imputed the correct set of potential outcomes for each individual in our study. This is known as a modus tollens argument,[56] and is the logical basis by which g-estimation is able to estimate parameters. Therefore, the estimate $\widehat{\psi}$ is taken to be the element of $\widetilde{\psi}$ at which abs$[Z(\widetilde{\psi})]$ is minimized.

There are three possible ways to obtain 95% confidence intervals for $\widehat{\psi}$. Slope-based estimated standard errors for $\widehat{\psi}$ can be computed as $\sigma_S = \left(d^{-2}\right)^{\frac{1}{2}}$, where d is the numerical derivative of the test statistic $Z(\widetilde{\psi})$ evaluated at $\widehat{\psi}$. This derivative may be estimated by the slope of a local linear regression of $\widetilde{\psi}$ on the $Z(\widetilde{\psi})$ statistic.[57] Alternatively, a bootstrap estimate of the standard error of $\widehat{\psi}$ can be obtained as the standard deviation of the set of 200 bootstrapped estimates $\left\{\widehat{\psi}^*_1, \cdots, \widehat{\psi}^*_{200}\right\}$. Each bootstrap estimate can be obtained by fitting model 2 to a random sample of size $N^* = 2{,}564$ (where * denotes a bootstrap resample) drawn, with replacement, from the original 2564 individuals.[20,58] Finally, assuming $Z\left(\widetilde{\psi}\right) \sim N(0,1)$, test-based upper and lower 95% confidence intervals for ψ can be computed by the values of $\widetilde{\psi}$ that correspond to $Z\left(\widetilde{\psi}\right) = \pm 1.96$.[59]

For our data, asbestos exposure for each year on study (X_t) followed a complex zero-inflated distribution. To simplify the analysis, we categorized this continuous exposure into 11 groups (one category for $X_t = 0$ and deciles for $X_t > 0$) and fit an ordinal logistic model[60] for Equation A2.1, which was implemented with PROC GENMOD, as:

$$\begin{aligned}
&\text{logit}\left[\frac{\pi_{m+1} + \cdots + \pi_{11}}{\pi_1 + \pi_2 + \cdots + \pi_m}\right] \\
&= \alpha_{0m} + \alpha_1 I(\text{nonwhite}) + \alpha_2 I(\text{female}) + \alpha_{f1} f_1(\text{year}) \\
&\quad + \alpha_{f2} f_2(\text{byr}) + \alpha_{f3} f_3(\text{age}) + \alpha_{f4} f_4(C) + \alpha_g g\left(X_{t-1}\right) \\
&\quad + \alpha_h \boldsymbol{I}^{(1)} + \alpha_t \boldsymbol{I}^{(2)} + \alpha_5 h[Y\left(\widetilde{\psi}\right)]\hat{\kappa}
\end{aligned}$$

Where π_j, j = 1 to 11 is the probability that the continuous exposure value falls in category j; $I(\cdot)$ is an indicator function; $f_k(\cdot)$, k = 1, 2, 3, 4 returns a 1×4 vector of restricted quadratic spline basis functions for argument $(\cdot)$ with knots as defined above; α_{fk} represents a 4×1 vector of parameters for each basis function; C is administrative censoring times (described in the next section) for each individual; $g(\cdot)$ returns a 1×10 vector of fractional polynomial functions for argument $(\cdot)$ of order $\pm 1/3, \pm 1/2, \pm 1, \pm 2, \pm 3$, $\boldsymbol{\alpha}_g$ represents a 10×1 vector of parameters for each polynomial term, and $h(\cdot)$ is a function of the potential outcomes $Y\left(\widetilde{\psi}\right)$, and inverse probability of censoring weights $\hat{k}$, all of which are described in subsequent sections. Furthermore, $\boldsymbol{I}^{(1)}$ and $\boldsymbol{I}^{(2)}$represent all first- and second-order interaction terms between the linear components of each term in the model (except $h[Y\left(\widetilde{\psi}\right)]\hat{\kappa}$), respectively. We used fractional polynomials instead of restricted quadratic splines to specify past exposure because of the large number of zero-valued knots obtained with the latter. Including a function of past exposure $g(X_{t-1})$ is required to obtain consistent exposure effect estimates using structural nested models.[17] Including the function of past exposure in our parametric accelerated failure-time models did not appreciably change the results. Parameters for this exposure model were estimated among person-years classified as having been at work using generalized estimating equations[61] with an independent working covariance matrix.[62]

ARTIFICIAL CENSORING

The potential outcomes imputed by model 2 are defined as a function of the observation time $Y = \min(T,C)$. In this section, we assume C is the administrative censoring time defined as the difference between the study's administrative end date and date of entry into the study. We handle censoring because of loss to follow up in the next section. If asbestos actually shortens survival, then the survival times of individuals with higher exposures are more likely to be observed in our study. By using Y instead of T to impute the potential survival time, we create an artificial association between exposure and the potential survival time $T(\widetilde{\psi})$, thus violating the assumption of no unmeasured confounding.[18] We use artificial censoring to account for this.[20,63]

To do so, we define a counterfactual censoring time $C\left(\widetilde{\psi}\right) = C\exp\left\{\widetilde{\psi} \times \min\left[A_t\right]\right\}$ if $\widetilde{\psi} \geq 0$, and $C\left(\widetilde{\psi}\right) = C\exp\left\{\widetilde{\psi} \times \max\left[A_t\right]\right\}$ if $\widetilde{\psi} < 0$, where $\min[A_t]$ and $\max[A_t]$ refer to the global empirical min and max exposure values (ie, the min and max values across all person-years in the South Carolina cohort), and C is the administrative censoring time. In our study, we did not perform a line search over the space defined by $\widetilde{\psi} < 0$, and thus use only the global empirical max exposure values. We then define a new potential survival time $T^*\left(\widetilde{\psi}\right) = \min\left\{T\left(\widetilde{\psi}\right), C\left(\widetilde{\psi}\right)\right\}$. This new variable $T^*\left(\widetilde{\psi}\right)$ is the potential survival time only for those individuals whose potential survival times would have been observed under any possible cumulative exposure regime $\bar{a}$ that falls within the global empirical min and max bounds. For example, if the relation between an individual's potential survival time and observed exposure (as defined by model 2) is such that they

would not have been observed under any exposure history that could have occurred in our study (defined by the global minimum and maximum exposure values), $T^*\left(\widetilde{\psi}\right)$ takes on the potential censoring time, as opposed to the potential survival time value, and is treated as a censored observation.

Because $T^*\left(\widetilde{\psi}\right)$ is a function of the administrative censoring time C, we include C in Equation A2.1 defined above.[2,64] Finally, in the previous section, we referred to $h(\cdot)$ as the functional form of the potential outcomes in our exposure model, which we define here as $h\left[T^*\left(\widetilde{\psi}\right)\right] = I\left[T\left(\widetilde{\psi}\right) \le C\left(\widetilde{\psi}\right)\right]$. Such a form for the potential outcomes has been shown to affect the efficiency, but not the consistency, of the estimator for ψ.[2,64] In our study, the proportion of events observed after artificial censoring over the range of the parameter space defined by $\widetilde{\psi}$ was between 3.3% (n = 86, $\widetilde{\psi} = 2.0$) and 5.5% (n = 142, $\widetilde{\psi} = 0$, ie, the actual proportion of failures), with a mean value of 4.9% $\left(\bar{n} = 106.46\right)$.

INVERSE PROBABILITY OF CENSORING WEIGHTS

As with standard parametric accelerated failure-time models, we assume censoring because of deaths from other causes or loss to follow-up is independent of lung cancer mortality conditional on measured covariates. We use inverse probability of censoring weights to account for right censoring because of drop out or competing risks when fitting a structural nested accelerated failure-time model.[65] For each individual at time t, define:

$$\hat{\kappa}_t = \prod_{j=0}^{t} \frac{\hat{P}\left(C_j = 0\right) \times (1 - C_j)}{\hat{P}(C_j = 0 \mid \boldsymbol{V}_j)}$$

where $\hat{P}\left(C_j = 0 | \boldsymbol{V}_j\right) = 1 / \left\{1 + \exp\left[-\left(\hat{\beta}_{0j} + \hat{\beta}_1 I(\text{nonCaucasian} + \hat{\beta}_2 I(\text{female}) + \hat{\beta}_3' f_1(\text{year}) + \hat{\beta}_4' f_2(\text{byr}) + \hat{\beta}_5' f_3(\text{age}) + \hat{\beta}_6' f_4(\text{base}))\right]\right\}$, and $\hat{P}\left(C_j = 0\right) = 1 / \left\{1 + \exp\left[-\left(\hat{\alpha}_{0j}\right)\right]\right\}$ are based on pooled logistic regression models with parameters estimated using maximum likelihood. Thus, if a person is uncensored in a given year j, $\hat{\kappa}_j$ is a function of the inverse of the conditional probability of being censored, and is zero otherwise. This inverse probability weight is stabilized by the marginal probability of being censored in year j. For our data, these stabilized weights were not extreme in their distribution, with a mean = 1.00, min = 0.54, and max = 16.59. As denoted in our model for Equation A2.1, we multiply $h\left[T^*\left(\widetilde{\psi}\right)\right]$ by $\hat{\kappa}$ to account for nonadministrative censoring. We refer the reader to Witteman et al[59(p 393)] for an explanation of how these inverse probability weights account for nonadministrative censoring using g-estimation.