eAppendix

G-estimation of accelerated failure-time models: A blueprint

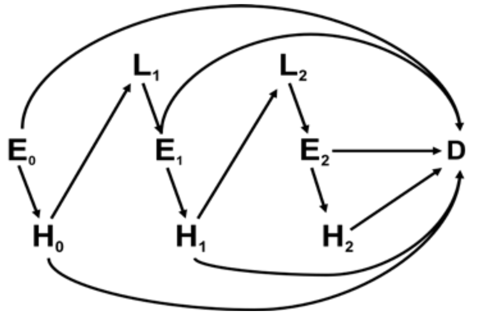

Figure A.1. Simplified DAG for the HWSE

Causal models have evolved substantially over the past decade and are becoming part of the mainstream epidemiologic literature.[1] These models aim to simulate randomized-controlled trials, contrasting outcomes that would have been observed in scenarios where the same individuals are subjected to different levels of exposure. Unlike traditional (conditional) models, causal models may be used to obtain unbiased parameter estimates when time-varying confounders, or their correlates, are situated on the causal path from exposure to outcome. This will occur if confounders are affected by prior exposure. Such a situation may arise under the healthy worker survivor effect (HWSE) illustrated by Figure A.1, which shows a simplified directed acyclic graph (DAG)[2] where arrows indicate causal relationships between outcome (D), and exposure (E), unmeasured health status (H) and time off work (L) at times 0, 1 and 2. This graph (which may not show all possible causal relationships in order to highlight those relevant to our discussion) assumes that exposure (E) affects health status (H), which itself affects both mortality risk (D) and exposure in subsequent years through increased time off work (L). L could also represent transfer to a job with lower exposure, use of protective equipment or employment termination. In the scenario described above, health status at time 0 (H_0) would thus act both as an intermediate on the causal pathway between exposure at time 0 (E_0) and outcome (D), and a confounder of the relation between exposure at time 1 (E_1) and outcome (D). The same applies to subsequent time points.

While traditional conditional methods yield biased estimates in situations such as the one described above, causal methods provide valid estimates by modeling exposure at every time point conditional on prior exposure and measured covariates (such as time off work). In occupational cohorts, however, exposure does not vary among unemployed individuals (they are always unexposed). This violates a key assumption needed to apply marginal structural models (the so-called positivity or experimental treatment assignment assumption), which requires that both exposed and unexposed individuals be observed within every covariate stratum.[3] G-estimation, in contrast, does not require this assumption and may thus be used to control for HWSE.[4]

G-estimation is a two-step iterative process whose objective is to unlink survival potential from observed exposure. Survival potential, in the form of survival time when never exposed [denoted $T(\psi)$], is first estimated using an accelerated failure-time (AFT) model that is based on an unknown parameter to which a candidate value is assigned. In the second step, the parameter value is tested by assessing the association between $T(\psi)$ and observed exposure $A(t)$ at each time *t*. These two steps, described in detail below, are repeated with the objective of finding the parameter value resulting in a $T(\psi)$ that is unrelated to $A(t)$. This "correct" $T(\psi)$ is denoted $T_{\bar{0}}$.

Step 1 - Estimating $T_{\bar{0}}$. The AFT model used to estimate $T_{\bar{0}}$ is shown below [equation (A.1)]. It relies on the assumption that every individual possesses a set of survival times that correspond to specific exposure histories, only one of which is observed. Since most of these survival times have not actually occurred, they are generally termed counterfactuals (i.e. contrary to fact). The model, which is a function of observed survival time ($T_{\bar{A}}$), $A(t)$ and an unknown parameter (ψ), assumes that every year of exposure affects survival time by $1 - e^{\psi}$ years. In the first step, a candidate value of ψ is selected and $T(\psi)$ is calculated for each individual. When the true value of the parameter is used, the function represents survival times that would have been observed had individuals never been exposed ($T_{\bar{0}}$).

Equation (A.1)

$$T(\psi) = \int_{0}^{T_{\bar{A}}} \exp\,[\psi A(t)]\,\mathrm{dt}$$

In our application, $A(t) = 1$ if workers were ever exposed to straight MWF in year t, and 0 otherwise. An individual with $T_{\bar{A}}$ = 10 years and 6 years of exposure would thus be determined to have $T(\psi)$= $6\exp(\psi) + 4$.

Step 2. G-estimation. In this step, g-estimation is used to identify the value of ψ that unlinks $T(\psi)$ from *A(t)* conditional on fixed covariates (L) and time-varying covariates (such as exposure history $\overline{A}$, history of time off work $\overline{O}$ and calendar year) up to time $t - 1$ in a pooled logistic regression model such as the one shown in equation (A.2). This analysis is performed with a dataset in which every row represents a person-year. Note that the analysis is restricted to individuals alive ($T_{\bar{A}}$ > t) and actively employed (*Em(t)* = 1) at time *t* and that $T(\psi)$ is expressed as a fixed (non-time-varying) variable.

Equation (A.2)

$$\begin{aligned} &logit\ Pr\left[A(t)|T(\psi), L, \overline{A}(t-1), \overline{O}(t-1), Em(t) = 1, T_{\bar{A}} > t\ \right] \\ &= \beta_0 + \beta_1 T(\psi) + \beta_2 A(t-1) + \beta_3 A(t-2)... + \beta_4 O(t-1) + \beta_5 O(t-2)... + \beta_6 L \end{aligned}$$

Steps 1 and 2 are repeated using an algorithm that performs a grid search until the value of ψ that unlinks $T(\psi)$ from *A(t)* (i.e. that results in $\beta_1 = 0$ and p-value = 1 in the pooled logistic model) is found. This value is denoted ψ^*. A manual grid search may also be implemented. Alternatively, non-gradient-based optimizers may be used.[5] Time-varying covariates may be expressed in various ways. We expressed exposure and off work histories as separate variables for $t - 1$, $t - 2$, $t - 3\ to\ 7$, $t - 8\ to\ 12$, and $t \geq 13$. These variables may however be expressed as a cumulative sum, an average or other measure.

Assuming no unmeasured confounders and correct model specification, $\psi^* = 0$ represents the null hypothesis, negative values represent beneficial exposures and positive values correspond to harmful exposures. The value $e^{-\psi^*}$ is the ratio of median survival times under a scenario in which the entire cohort is exposed without interruption throughout follow-up ($T_{\bar{1}}$) to that in which the entire cohort is never exposed ($T_{\bar{0}}$).

Equation (A.3)

$$\frac{T_{\bar{1}}}{T_{\bar{0}}} = e^{-\psi^*}$$

In our study, as well as in many occupational settings, very few workers were exposed to MWF without interruption. Equation (A.3) may be generalized to estimate the contrast between the median survival time if the entire cohort had been exposed to straight MWF for the first m years ($T_{\bar{a}}$) relative to the median survival time had the cohort remained unexposed throughout follow up ($T_{\bar{0}}$).

Equation (A.4)

$$\frac{T_{\bar{a}}}{T_{\bar{0}}} = \frac{T_{\bar{0}} + m[1 - \exp(\psi)]}{T_{\bar{0}}}$$

We determined 95% confidence intervals by bootstrapping the above quantity 200 times. Alternatively, the lower and upper bounds of 95% confidence intervals may be identified by repeating the procedure described above with the objective of finding the two values of ψ that result in p-values of 0.025 for β_1 using a robust variance estimator (e.g. the Huber-White sandwich estimator) that considers the non-independent structure of the data. Other methods, based on the Wald statistic and a variance estimate for ψ^*,[6] may also be used.

Competing Risks and Loss to Follow-Up

Because exposure and health status may result in loss to follow-up and/or deaths from causes other than the one under study (also referred to as competing risks), these types of censoring may result in selection bias.[7] Assuming that all variables predicting mortality by competing risk and loss to follow-up were measured, one may correct for this bias by inverse probability weighting (IPW). This method assigns weights to uncensored individuals that are equal to the inverse of the probability of being uncensored at time *t*. In this manner, a person with a probability of not dying of competing cause and of not being lost to follow-up equal to 20 percent would be assigned a weight of five since four other individuals with the same characteristics would have been included in the analysis had censoring not occurred. Weights are determined by logistic regression with the dependent variable being a dichotomous time-varying variable coded 1 if individuals were uncensored and 0 if they were censored at time *t*. Independent variables include fixed and time-varying covariates, such as history of exposure and time off work up to $t - 1$. For each individual, a running product of the predicted probabilities yields the cumulative probability of remaining uncensored at each time *t*; the inverse of this product is used as a weight in equation (A.2). Individuals who are censored by competing risks or loss to follow-up are excluded from this analysis. The distribution of the weights should be examined and consideration should be given to truncating (i.e. replacing with smaller values) any extreme weights that may inflate confidence intervals.[8] No extreme weights were observed in our analysis (range: 1.06-1.74) and so no truncation was performed.

Administrative Censoring

Unexposed survival time cannot be computed using the accelerated failure-time model in equation (A.2) for individuals who survived beyond the end of follow up (i.e. were administratively censored), and restricting analyses to those whose outcome was observed may introduce selection bias. Restricting analyses to individuals whose outcome would have been observed under all exposure scenarios has however been shown to eliminate this bias.[9] For individuals exposed to a harmful substance (i.e. if ψ is positive), being never exposed ($T_{\bar{0}}$) would result in the longest survival time. In this case, analyses should therefore be restricted to individuals whose outcome would have been observed even if they had never been exposed, i.e. those for whom $T_{\bar{0}} < C$, where C is the amount of time between the individual's enrollment in the study and the date of end of follow-up. Conversely, survival when always exposed would result in the longest survival time for a beneficial exposure (i.e. if ψ is negative) and thus analyses would be restricted to those whose outcome would have been observed had they been exposed throughout the study period. Adjustment for administrative censoring is thus achieved by conditioning on (i.e. adjusting for) C and replacing $T(\psi)$ by $\Delta(\psi)$ in equation (A.2), where for each individual $\Delta(\psi)$ is equal to 1 if the outcome of interest would have been observed and 0 otherwise. The value of $\Delta(\psi)$ is thus determined using the following equations:

For positive candidate values of ψ:

Equation (A.5)

$$\Delta(\psi) = 1 \text{ if } T(\psi) < C, \text{ and}$$
$$\Delta(\psi) = 0 \text{ if } T(\psi) \geq C$$

For negative candidate values of ψ:

Equation (A.6)

$$\Delta(\psi) = 1 \text{ if } \mathrm{T}(\psi)e^{-\psi} < C, \text{ and}$$
$$\Delta(\psi) = 0 \text{ if } \mathrm{T}(\psi)e^{-\psi} \geq C$$

Note that another function of C and $\mathrm{T}(\psi)$ has been shown to result in narrower confidence intervals than the one shown above[10] but is more computationally challenging.

Transforming survival ratios into hazard ratios

Although, as described in the Discussion section, survival ratios may be of greater public health relevance than hazard ratios, the epidemiological literature more commonly reports hazard ratios and so a method to transform the former into the latter may be of interest. At least two methods may be used to generate hazard ratios. First, a standard Cox proportional hazards model may be run on counterfactual survival and censoring times. In order to do so, a dataset with survival times under both unexposed ($T_{\bar{0}}$) and exposed ($T_{\bar{a}}$) scenarios must be generated. $T_{\bar{0}}$ is obtained using equation (A.1) and $T_{\bar{a}}$ may be computed using the numerator of equation (A.4) where:

Equation (A.7)
$$T_{\bar{a}} = T_{\bar{0}} + m[1 - \exp(\psi)]$$

Under the unexposed scenario, individuals who are administratively censored (i.e. with $\Delta(\psi) = 0$) are assumed to have been censored at time C if ψ is positive and at time Ce^{ψ} if ψ is negative. Censoring times under the exposed scenario $\bar{a}$ are computed using the censoring times under no exposure in equation A.7. Hazard ratios may also be computed nonparametrically. For both methods, confidence intervals may be determined by bootstrapping, and censoring by competing risk or loss to follow up is handled using IPW as described above.

Summary

We presented the procedure in logical order for pedagogical reasons. In practice, the steps for implementing g-estimation will be performed in the following order.

1. Fit a logistic regression model to estimate the probability of being free from competing risk and not lost to follow-up; calculate the IPWs.

2. Choose a candidate value for ψ and calculate $T(\psi)$, the candidate value for survival time under no exposure, based on the accelerated failure time model shown in equation (A.1).

3. Determine $\Delta(\psi)$ based on equations (A.5) or (A.6).

4. Using the weights determined in step 1, fit a weighted pooled logistic regression to estimate whether β_1= 0 in equation (A.2) but replace $\mathrm{T}(\psi)$ by $\Delta(\psi)$ and adjust for *C*.

5. Repeat steps 2-4 until the value of ψ that results in β_1= 0 (or as close to 0 as possible) is found.

6. Determine the upper and lower bounds of 95% confidence intervals. This may be done either by bootstrapping steps 2-5, by repeating these steps to find values of ψ that result in p-values of 0.025 for β_1, or based on a Wald test.

7. Determine the survival ratio using equation (A.3) to compare scenarios in which the entire population is exposed throughout follow-up to that in which the population is never exposed, or equation (A.4) to obtain the effect of m years of exposure relative to no exposure.

8. If desired, estimate the hazard ratio by a) generating a dataset comprising survival and censoring times under both unexposed and exposed counterfactual scenarios using equations (A.1) and (A.7), respectively, and b) running an unadjusted Cox model with exposure scenario as the independent variable. Confidence intervals for the hazard ratio may be determined by bootstrapping steps 2-5.

Note that g-estimation may also be performed by minimizing an estimating equation.[5] This method, while being more conceptually challenging, may require less computing time. We thus elected to use the estimating equation in the present paper and in STATA code available on our web site (http://ehs.sph.berkeley.edu/eisen/index.html).

References

1. Hogan JW. Bringing causal models into the mainstream. Epidemiology 2009;20:431-2.

2. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;10:37-48.

3. Robins JM, Hernán MA, Rotnitzky A. Effect modification by time-varying covariates. Am J Epidemiol 2007;166:994-1002; discussion 3-4.

4. Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures. In: Longitudinal data analysis. New York, NY: Chapman & Hall; 2009:553-99.

5. 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;14:477-91.

6. Robins JM, Tsiatis AA. Correcting for Noncompliance in Randomized Trials Using Rank Preserving Structural Failure Time Models. Commun Stat Theory 1991;20:2609-31.

7. Hernán MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004;15:615-25.

8. Kish L. Weighting for Unequal Pi. J Official Stat 1992;8:183-200.

9. 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;12:1605-28.

10. 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. Epidemiology 1992;3:319-36.