

## G-estimation of Causal Effects: Isolated Systolic Hypertension and Cardiovascular Death in the Framingham Heart Study

Jacqueline C. M. Witteman,<sup>1</sup> Ralph B. D'Agostino,<sup>2</sup> Theo Stijnen,<sup>1</sup> William B. Kannel,<sup>3</sup> Janet C. Cobb,<sup>2</sup> Maria A. J. de Ridder,<sup>1</sup> Albert Hofman,<sup>1</sup> and James M. Robins<sup>4</sup>

Time-dependent covariates are often both confounders and intermediate variables. In the presence of such covariates, standard approaches for adjustment for confounding are biased. The method of G-estimation allows for appropriate adjustment. Previous studies applying the G-estimation method have addressed effects on all-cause mortality rather than on specific causes of death. In the present study, a method to adjust for censoring by competing risks is presented. The authors used the approach to estimate the causal effect of isolated systolic hypertension on cardiovascular mortality in the Framingham Heart Study, with a 10-year follow-up using data from 1956 to 1970. Arterial rigidity is a major determinant of isolated systolic hypertension and may be a confounder of the relation between isolated systolic hypertension and cardiovascular death. Conversely, isolated systolic hypertension may by itself contribute to stiffening of the vessel wall, and arterial rigidity may therefore also be an intermediate variable in the causal pathway from isolated systolic hypertension to cardiovascular death. While controlling for arterial rigidity and other baseline and time-dependent covariates, isolated systolic hypertension decreased the time to cardiovascular death by 45% (95% confidence interval 3–69). *Am J Epidemiol* 1998;148:390–401.

bias (epidemiology); blood pressure; cardiovascular diseases; epidemiologic methods; follow-up studies; models, statistical; statistics

Isolated systolic hypertension (ISH) is an established risk factor for cardiovascular disease (1). However, arterial stiffening is the prime cause of ISH and may be associated with an increased cardiovascular risk (2, 3). That means that arterial rigidity may be a confounder of the observed relation between ISH and cardiovascular disease (4, 5). Conversely, high arterial pressure may cause structural changes in the arterial wall and thereby contribute to arterial rigidity (6–8). Arterial rigidity may therefore also be an intermediate variable on the causal pathway from ISH to cardiovascular disease. If this is the case, a Cox proportional hazards model that includes the time-dependent covariate arterial rigidity will underestimate the overall

effect of ISH on cardiovascular disease. This is a well-recognized problem with time-dependent Cox proportional hazards models (9). The method of G-estimation of structural failure time models is an alternative approach that allows for appropriate adjustment of the effect of an exposure in the presence of time-varying risk factors for death that affect future exposure and are themselves influenced by past exposure (10–17). We used this method to estimate the causal effect of ISH on cardiovascular death in the Framingham Heart Study. Previous studies applying the G-estimation method have addressed effects on all-cause mortality rather than on specific causes of death. In this study, we include a method to adjust for censoring by competing risks. The aim of the paper is twofold: 1) to discuss the main characteristics of the method and 2) to apply the method to data from the Framingham Heart Study. A detailed overview of G-estimation of structural failure time models is provided in reference 17.

Received for publication June 17, 1996, and accepted for publication February 4, 1998.

Abbreviations: GEE, generalized estimating equations; ISH, isolated systolic hypertension.

<sup>1</sup> Department of Epidemiology and Biostatistics, Erasmus University Medical School, Rotterdam, Netherlands.

<sup>2</sup> Department of Mathematics, Boston University College of Liberal Arts, Boston, MA.

<sup>3</sup> Section of Preventive Medicine and Epidemiology, Boston University School of Medicine, Boston, MA.

<sup>4</sup> Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA.

Reprint requests to Dr. J. C. M. Witteman, Department of Epidemiology and Biostatistics, Erasmus University Medical School, P.O. Box 1738, 3000 DR Rotterdam, Netherlands.

### MATERIALS AND METHODS

#### The Framingham Heart Study

The Framingham cohort of 2,336 men and 2,873 women, aged 30–62 years at entry, has been exam-

ined biennially since 1948. The sampling procedure, design, and methods of the study have been presented elsewhere (18). Data collected at biennial visits include systolic and diastolic blood pressure, use of antihypertensive medication, serum total cholesterol, cigarette smoking, metropolitan relative weight, and diabetes mellitus. We define ISH as a systolic pressure of 160 mmHg or higher, with a diastolic pressure less than 95 mmHg. Diastolic hypertension is defined as a diastolic blood pressure of 95 mmHg or more independent of the level of systolic blood pressure. We use the presence of aortic calcified plaques, as detected on chest radiographs, as a measure of arterial rigidity. Calcified plaques in the aorta are known to represent atherosclerosis (19), which is a cause of rigidity of the aorta (2, 7). Previous studies have shown that aortic calcification is a predictor of cardiovascular morbidity and mortality (20, 21). Roentgenograms of the thorax were routinely taken at biennial examinations. Calcified plaques were diagnosed as present when typically shaped densities were seen in the aortic arch or in the descending part of the thoracic aorta. The scoring was performed at visits 5–10. Aortic calcified plaques were considered present after the first manifestation, even though the plaques might not be visualized on all of the subsequent visits. Participants were monitored for development of cardiovascular disease during follow-up. The endpoint that is considered in this study is death from cardiovascular disease (primarily myocardial infarction, stroke, and sudden coronary death).

#### Data description and notation

Exposure status is defined by the presence or absence of ISH and may alternate between visits. Potential confounders considered in this study are cigarette smoking, serum total cholesterol, metropolitan relative weight, diabetes mellitus, and aortic calcification. Confounders measured at or before baseline will be referred to as baseline covariates. Time-dependent levels of the potential confounders will be referred to as time-dependent covariates. Data from visits 5–11 were made available to us. We defined the at risk period (i.e., the follow-up period) to be the period from visits 6 to 11 so the variables available at visits 5 and 6 could be used to adjust for confounding present at the start of the follow-up. The cohort consists of 4,404 subjects alive and with available exposure and covariate information at visit 6. For subjects with missing values at a follow-up visit but with measurements present at subsequent visits, we replaced the missing value with the value from the closest previous visit. During 10 years of follow-up (visits 6–11), 306 subjects died from cardiovascular disease. Visits are denoted by  $k$ .  $ISH_{i,k}$  is the ISH status of subject  $i$  at visit  $k$ . A similar

notation is used for the covariates. For any time-dependent variable, we use overbars to denote the history of that variable up to and including  $k$ . For example,  $\overline{ISH}_{i,k} = (ISH_{i,k}, ISH_{i,k-1}, ISH_{i,k-2}, \text{ and so on})$ . We assume that ISH status at  $k$  remains unchanged for the interval  $(k, k + 1)$ . We use  $ISH_i(t)$  to refer to ISH status at some time  $t$  between visits. With  $\overline{ISH}_i(t)$ , we refer to the history of ISH up to  $t$ . A subject's failure time is denoted by  $T_i$ , measured in visit units since visit 6. One visit unit equals 2 calendar years.

#### The method of G-estimation

To define the causal effect of an exposure, we assume the presence of counterfactual failure times. Following Robins and coauthors (10–17), we assume that for each individual there exists a set of unobserved failure times, each corresponding to a defined exposure history. These failure times are counterfactual because they represent outcomes under circumstances that may not have actually occurred. A causal effect is defined in terms of a comparison of counterfactual failure times associated with different exposure histories. Specifically, we are interested in the comparison of the counterfactual failure times corresponding to the exposure history always exposed and the counterfactual failure times corresponding to the exposure history never exposed.

We define the counterfactual failure time  $U_i$  to be the time to failure if subject  $i$  is never exposed. In our study, never exposed means never having had ISH. In the next section we will describe a model that relates  $U_i$  to observed exposure history and observed failure time. Subsequently, we show how we estimate the model parameter and how this parameter relates to our causal parameter of interest. Counterfactual failure times can only be computed for subjects with an observed failure time. For pedagogic reasons, we first discuss the G-estimation method in the absence of censoring. This would apply to studies in which the outcome of interest is all causes of death and the cohort is followed to extinction.

*A model for the counterfactual failure time.* We define a failure time model that relates a person's counterfactual failure time,  $U_i$ , to his observed exposure history and observed failure time, as proposed by Cox and Oakes (22) and by Robins and coauthors (10–17). In the present study,  $U_i$  is the time to failure if individual  $i$  never had ISH. Like  $T_i$ ,  $U_i$  is measured in visit units since visit 6. A model that relates  $U_i$  to the observable data  $(T_i, \overline{ISH}_i(T_i))$  is

$$U_i = \int_6^{6+T_i} \exp\{\psi_0 * \overline{ISH}_i(t)\} dt, \quad (1)$$

where  $\psi_0$  is an unknown parameter that we will later estimate.  $U_i$  is a weighted sum of time spent in a given ISH status, where the weights are  $\exp(\psi_0)$  if  $ISH_i = 1$  and 1 if  $ISH_i = 0$ . For example, if  $\psi_0 = 0.5$  and the observed data for subject  $i$  are  $T_i = 2.8$  visit units (5.6 years),  $ISH_{i,6} = 0$ ,  $ISH_{i,7} = 1$ , and  $ISH_{i,8} = 1$ , then  $U_i = (7 - 6)\exp(0.5 * 0) + (8 - 7)\exp(0.5 * 1) + (8.8 - 8)\exp(0.5 * 1) = 4.0$  (8 years). Thus, if this subject never had ISH, his time to failure would have been 8 instead of the observed 5.6 years.

When  $\psi_0 = 0$ , equation 1 becomes  $U_i = T_i$ , for all  $i$  (i.e., each subject's counterfactual failure time  $U_i$  equals his observed failure time,  $T_i$ ). Hence,  $\psi_0 = 0$  corresponds to the null hypothesis that exposure has no effect. To understand the implications of the model when  $\psi_0$  is not 0, we have to consider another counterfactual failure time, the failure time that corresponds to continuous exposure. Our model implies that this counterfactual failure time can be obtained by multiplying the counterfactual failure time if never exposed ( $U_i$ ) by the factor  $\exp(-\psi_0)$ . That is, a subject's unexposed lifetime  $U_i$  is expanded or contracted by the factor  $\exp(-\psi_0)$  because of continuous exposure. If  $\psi_0 > 0$ , exposure decreases survival, and if  $\psi_0 < 0$ , exposure increases survival. In previous papers, this model is called the strong version of the time-dependent accelerated failure time model or the rank-preserving structural failure time model (15, 22).

**Estimation of the exposure effect.** In this section, we discuss how we estimate the true value of  $\psi_0$ . Fundamental to the approach is the assumption of no unmeasured confounders. We say that there are no unmeasured confounders when, conditional on past ISH and recorded covariate history, subjects with and subjects without ISH at visit  $k$  would have the same survival if, possibly contrary to fact, they had never been exposed. Consequently, under the assumption of no unmeasured confounders, ISH at  $k$  is independent of  $U_i$  given the recorded past. To test this independence, we postulate a model for the probability that an individual has ISH at  $k$  given past ISH and covariate history. Define

$$U_i(\psi) = \int_6^{6+T_i} \exp\{\psi * ISH_i(t)\} dt, \quad (2)$$

so that by equation 1,  $U_i(\psi) = U_i$  if  $\psi = \psi_0$ . For each value of  $\psi$ ,  $U_i(\psi)$  can be computed from the data ( $T_i$ ,  $ISH_i(T_i)$ ). If equation 1 is correctly specified, and assuming no unmeasured confounders, if  $\psi$  equals the true value  $\psi_0$ , then the true value of  $\theta$  in the following

logistic model equals 0: for  $k = 6, 7, \dots, 10$ ,

$$\text{logit } P(ISH_{i,k} = 1 | \overline{ISH}_{i,k-1}, \overline{\mathbf{L}}_{i,k}, U_i(\psi), T_i > k) = \alpha_k + \beta \mathbf{W}_{i,k} + \theta U_i(\psi). \quad (3)$$

Here,  $\overline{\mathbf{L}}_{i,k}$  is a vector consisting of all baseline and time-dependent covariates up to and including  $k$ ,  $\mathbf{W}_{i,k}$  is a vector of functions of  $(\overline{ISH}_{i,k-1}, \overline{\mathbf{L}}_{i,k})$ , and  $\alpha_k$ ,  $\beta$ , and  $\theta$  are unknown parameter vectors. We test the hypothesis that a particular value of  $\psi$  equals  $\psi_0$  by testing whether  $\theta = 0$  in equation 3 using a score, Wald statistic, or likelihood ratio test. We refer to any such test of the hypothesis  $\psi = \psi_0$  as a G-test. Our G-estimate  $\hat{\psi}$  of  $\psi_0$  is the value of  $\psi$  for which our G-test of the hypothesis  $\theta = 0$  is 0. Equivalently,  $\hat{\psi}$  is the value of  $\psi$  for which our G-test has a  $p$  value = 1. In fitting the model, we pool over all visits  $k$ ; that is, each individual contributes multiple observations, one for each time  $k$  the subject is alive. The variance  $\sigma^2$  of our point estimate,  $\hat{\psi}$ , can be consistently estimated by squaring the inverse of the estimate of the slope of our G-test statistic  $Z(\psi)$  evaluated at  $\hat{\psi}$  (23). A 95 percent confidence interval can be constructed by the Wald statistic, using the estimated variance,  $\sigma^2$ , or by a test-based method. The test-based 95 percent confidence interval for  $\psi_0$  consists of those values of  $\psi$  for which a G-test of  $\theta = 0$  fails to reject at a 5 percent level (13). Conceptually, the G-estimation procedure checks, at each successive time  $k$ , for an association between ISH status at  $k$  and the hypothesized value  $U_i(\psi)$  of the true but unknown counterfactual failure time  $U_i$ , 1) after adjusting for exposure and covariate history before  $k$ , but 2) without adjusting for covariate status subsequent to  $k$ . By separately examining ISH status at each successive time  $k$ , the G-estimation procedure succeeds in controlling confounding by intermediate variables.

**Censoring by end of follow-up.** In this section, we suppose there is no censoring by competing risks; that is, we consider all causes of death. However, we allow censoring by end of follow-up. We define the potential censoring time at end of follow-up,  $C_i$ , as the time from the baseline visit (visit 6) to a predefined end of follow-up time (visit 11). Thus, in our data,  $C_i$  equals five visit units for all subjects. In the above, the counterfactual failure time,  $U_i(\psi)$ , can only be computed for uncensored individuals. We shall therefore replace  $U_i(\psi)$  in equation 3 by a function of  $U_i(\psi)$  and  $C_i$  which is observed for all subjects. The rationale is that, if  $U_i$  is independent of  $ISH_{i,k}$  given past ISH and covariate history, the same should be true for any function of  $U_i$  and  $C_i$  (since  $C_i$  is a baseline covariate).

Let

$$\begin{aligned}\Delta_i(\psi) &= 1 && \text{if } U_i(\psi) < C_i(\psi) \text{ and} \\ \Delta_i(\psi) &= 0 && \text{if } U_i(\psi) \geq C_i(\psi),\end{aligned}\quad (4)$$

where  $C_i(\psi) = C_i$  if  $\psi \geq 0$  and  $C_i(\psi) = C_i \cdot \exp(\psi)$  if  $\psi < 0$ . With this definition of  $C_i(\psi)$ , we obtain that when a subject is censored ( $T_i > C_i$ ),  $U_i(\psi) > C_i(\psi)$ , so  $\Delta_i(\psi)$  is always observed. Furthermore,  $\Delta_i(\psi)$  is a function only of  $U_i(\psi)$  and  $C_i$ . When  $\Delta_i(\psi) = 0$ , we say an individual is  $\psi$ -censored. To generate our estimates, we would like to use a simple function with good power. The function  $\Delta_i(\psi)$  satisfies this criterion; that is, we will replace  $U_i(\psi)$  with 1 for individuals not  $\psi$ -censored and with 0 for those  $\psi$ -censored. Indeed, in our data, the use of  $\Delta_i(\psi)$  gave better power than the alternative choices  $X_i(\psi) = \min(U_i(\psi), C_i(\psi))$  or  $\Delta_i(\psi)X_i(\psi)$ .

**Censoring by competing risks.** In the previous section, we considered the case in which the only type of censoring is censoring by end of follow-up. We will now discuss how we deal with censoring by competing risks when our endpoint of interest is death from cardiovascular disease. Let us for the moment assume that the only cause of censoring other than end of follow-up is death from other causes. We assume that we have data on a sufficient number of confounding factors so that, conditional on these covariates, censoring due to competing risks (i.e., death from other causes) is independent of the time to cardiovascular death. We then specify a model for the censoring process. We specify a logistic model for the probability of death from other causes in the interval  $(k, k + 1)$ , given being alive at  $k$ , and with death from cardiovascular disease considered as a censoring event. The model is of the form: for  $k = 6, 7, \dots, 10$ ,

$$\begin{aligned}\text{logit } P(D_i \leq k + 1 | \overline{\text{ISH}}_{i,k}, \overline{\mathbf{L}}_{i,k}, D_i > k, T_i > k) \\ = \boldsymbol{\alpha}_k^* + \boldsymbol{\beta}^* \mathbf{W}_{i,k}^*.\end{aligned}\quad (5)$$

Here,  $T_i$  is redefined as time to death from cardiovascular disease,  $D_i$  is the time to death from other causes,  $\overline{\mathbf{L}}_{i,k}$  is a vector consisting of all baseline and time-dependent covariates up to and including  $k$ ,  $\mathbf{W}_{i,k}^*$  is a vector function of  $(\overline{\text{ISH}}_{i,k}, \overline{\mathbf{L}}_{i,k})$ , and  $\boldsymbol{\alpha}_k^*$  and  $\boldsymbol{\beta}^*$  are unknown parameter vectors. Because the probability of being censored by competing risks between  $k$  and  $k + 1$  is small, the logistic model will lead to parameter estimates close to those obtained by a continuous time Cox proportional hazard model (24). For each subject who does not suffer death from other causes, we use equation 5 to estimate the cumulative probability  $K_i$  of being free from competing risks through

the last visit (time to death from cardiovascular disease or end of follow-up). Our estimates,  $\hat{K}_i$  of the probability  $K_i$ , are obtained by multiplying the estimated conditional probabilities of being free from competing risks in each 2-year interval prior to a subject's last visit. We then further modify our G-estimation procedure by replacing  $\Delta_i(\psi)$  (which was substituted for  $U_i(\psi)$  in equation 3) by  $\Delta_i(\psi)/\hat{K}_i$  if a subject is uncensored by competing risks and by 0 if a subject is censored by competing risks. An intuitive explanation is given here. Given the correctness of our assumption of conditionally independent censoring by competing risks, the following will be true: for each person with a cumulative probability of 25 percent of being free from death by competing risks through the last visit, there would, on average, have been three other persons who were censored by competing risks before or at the last visit, but who would have had a similar set of covariates and risk of cardiovascular death, had censoring been prevented. We therefore assign this person a weight of four in the G-estimation by multiplying his covariate  $\Delta_i(\psi)$  by the factor four.

In the present study, we have to consider other causes of censoring by competing risks. Our goal is to learn what the effect of ISH is on death from cardiovascular disease when diastolic blood pressure in the population is normal and antihypertensive medication is not given. To do so, subjects who develop diastolic hypertension or start antihypertensive treatment must be regarded as censored by competing risks at the time they reach one of these conditions. Specifically, we fit a polytomous logistic regression model with death from other causes and the development of diastolic hypertension or start of antihypertensive treatment as outcome events to obtain estimates of the conditional probability of being free from all censoring events. We then compute the cumulative probability  $\hat{K}_i$  and modify our G-estimation procedure by weighting by the inverse of  $\hat{K}_i$ , as described above. Furthermore, we no longer regard subject  $i$  as an observation at visit  $k$  if subject  $i$  is censored prior to visit  $k$ , that is, subject  $i$  has developed diastolic hypertension or started antihypertensive treatment or died.

In the presence of competing risks, our G-estimate  $\hat{\psi}$  is still a consistent, asymptotically normal estimator. However, our confidence intervals are no longer valid because 1) the subject-specific contributions to the G-test statistic from two different visits are now correlated and 2) we have used an estimate  $\hat{K}_i$  rather than the true, but unknown,  $K_i$ . However, it follows from results on pages 284 and 285 of reference 14 and the Appendix of reference 17 that conservative 95 percent large-sample confidence intervals (i.e., intervals guaranteed to cover  $\psi_0$  at least 95 percent of the time) can

be obtained by 1) regarding the exposure variables  $ISH_{i,k}$  for a given subject  $i$  as clustered (i.e., correlated) binary data and 2) constructing a G-test of the hypothesis  $\theta = 0$  in equation 3 using the robust Wald test outputted by a generalized estimating equations (GEE) software package for clustered binary data using the independence working covariance matrix option (25). In the presence of censoring, one will usually obtain better and shorter confidence intervals by using the Wald rather than test-based intervals when the data set is relatively small. This is due to loss of information caused by artificial censoring for large positive or negative values of  $\psi$ .

**$\psi_0$  and the causal rate ratio.** The expansion factor  $\exp(-\psi_0)$  is related to a common parameter of public health interest. Specifically,  $1 - \exp(-\psi_0)$  is the fractional decrease in life due to continuous exposure. Greenland and Robins (26) have argued that measures of change in life expectancy are often of greater public health interest than rate ratio measures. However, because the rate ratio is commonly used in epidemiology, we will characterize the relation between  $\psi_0$  and the causal rate ratio. We define the causal rate ratio as the mortality rate if all subjects were exposed from baseline onward divided by the rate when all subjects were unexposed throughout. The relation depends on the shape of the distribution of  $U_i$ . If the underlying distribution of  $U_i$  is Weibull, that is,  $P(U_i > t) = \exp(-\{\lambda t\}^\kappa)$ , the causal rate ratio equals  $\exp(\kappa\psi_0)$ . For subjects who reach end of follow-up without failing ( $T_i > C_i$ ), we take  $C_i(\psi_0)$  as their time of censoring by end of follow-up, assuming independence of  $U_i$  and  $C_i$ . The parameter  $\kappa$  will be estimated using maximum likelihood. Subjects who are censored by competing risks are not included in the estimation procedure. However, information from subjects censored by competing risks will be used by weighting the estimation procedure; the contribution to the log likelihood of each included subject  $i$  is weighted by the inverse of the probability,  $\hat{K}_i$ , of being free from competing risks.

## RESULTS

### Evaluation of time-dependent confounders

We first examined the data for time-dependent confounders. We have time-dependent confounders if 1) a subset of the time-dependent covariates are predictors of the hazard of death by cardiovascular disease at any time  $t$ , controlling for past ISH history and baseline covariates, and 2) a subset of the time-dependent covariates predict ISH at some visit  $k$ , given past ISH history and baseline covariates. We used the time-dependent Cox proportional hazards model to examine whether covariate status before  $t$  was related to the

hazard of cardiovascular death at  $t$ , conditional on past ISH and baseline covariate status. Cigarette smoking, serum cholesterol, metropolitan relative weight, diabetes mellitus, and aortic calcification were evaluated as time-dependent covariates. The main predictors of cardiovascular death are given in table 1. The risk of death was significantly related to diabetes mellitus at the preceding visit (visit  $k$ ). Diabetes mellitus at visits before  $k$  was not a significant predictor of death after adjustment for diabetes mellitus at  $k$ . The association of cardiovascular death with aortic calcification increased with increasing time period between death and first occurrence of calcification. Borderline significant associations of cardiovascular death with aortic calcification at visits before  $k$  were still present after adjustment for more recent measurements. No significant associations were observed between cardiovascular death and other time-dependent covariates.

We used a logistic model, pooling over all visits  $k$  (i.e., visits 6–10), to examine whether time-dependent covariates predicted ISH at  $k$ , conditional on past ISH and baseline covariate status. We assumed that only levels of covariates measured through  $k - 1$  were prior to the assessment of ISH at  $k$ , and so covariates at visit 5 were treated as baseline covariates. Subjects were censored when they developed

**TABLE 1. Relative risks of cardiovascular death,\* Framingham Heart Study, 10-year follow-up using data from 1956 to 1970**

	RR†	95% CI‡
Age (per year)	1.09	1.08–1.10
Sex (male = 0, female = 1)‡	0.55	0.44–0.69
Baseline cigarette smoking (per 10 cigarettes/day)§	1.21	1.12–1.35
Baseline serum cholesterol (per 10 mg/dl)§	1.02	1.0–1.05
Baseline metropolitan relative weight (per 10%)§	1.05	1.0–1.12
Baseline diabetes mellitus§	3.43	2.38–4.92
Diabetes mellitus at $k$	1.96	1.25–3.07
Baseline aortic calcification§	1.53	1.19–1.98
Aortic calcification at $k$	1.11	0.83–1.49
Aortic calcification at $k - 1$	1.21	0.88–1.65
Aortic calcification at $k - 2$	1.40	0.99–1.99
Aortic calcification at $k - 3$	1.44	0.94–2.19
Aortic calcification at $k - 4$	1.81	1.01–3.22

\*Using the time-dependent Cox proportional hazards model.

† RR, relative risk; CI, confidence interval.

‡ Adjusted for age.

§ The model (baseline model) includes age, sex, isolated systolic hypertension at visit 5 ( $ISH_5$ ) and at visit  $k$  ( $ISH_k$ ), and all baseline covariates at visit 5 (cigarette smoking, serum cholesterol, metropolitan relative weight, diabetes mellitus, and aortic calcification).  $k$  is the last visit on the current time axis.

|| The model includes all variables of the baseline model and the time-dependent covariate. Models including aortic calcification at visits before  $k - 1$  include only part of the follow-up period because of missing covariate status.  $k$  is the last visit on the current time axis.

diastolic hypertension or started antihypertensive treatment. The main predictors of ISH are given in table 2. Aortic calcification at  $k - 1$  was a predictor of  $ISH_k$ . Aortic calcification at previous visits was not significantly associated with  $ISH_k$  after adjustment for the most recent measurement preceding ISH. Diabetes mellitus at  $k - 1$  had a weak, nonsignificant association with  $ISH_k$ . We next examined whether the occurrence of aortic calcification was predicted by ISH. We used a pooled logistic model to assess the probability of aortic calcification at  $k$  as a function of baseline ISH, conditional on baseline covariates, absence of diastolic hypertension or antihypertensive treatment at baseline, and absence of aortic calcification at  $k - 1$ . The relative risk associated with baseline ISH was 1.25 (95 percent confidence interval 0.98–1.60). Thus, the time-dependent covariate aortic calcification is an independent risk factor for cardiovascular death, predicts future ISH status, and is predicted by previous status of ISH.

### Modeling of censoring by competing risks

We next modeled censoring by competing risks. During follow-up, 277 subjects died from causes other than cardiovascular disease. Diastolic hypertension or use of antihypertensive treatment was present in 1,146 subjects at baseline and occurred in 590 during follow-up. We modeled the probability of death from other causes and the probability of the development of diastolic hypertension or start of antihypertensive treatment in the interval  $(k, k + 1)$  as functions of visit

**TABLE 2. Relative risks of isolated systolic hypertension (ISH),\* Framingham Heart Study, 10-year follow-up using data from 1956 to 1970**

	RR†	95% CI‡
Age (per year)‡	1.08	1.07–1.09
Sex (male = 0, female = 1)§	1.18	1.01–1.38
Baseline cigarette smoking (per 10 cigarettes/day)¶	1.11	1.04–1.19
Baseline serum cholesterol (per 10 mg/dl)¶	1.02	1.0–1.04
Baseline metropolitan relative weight (per 10%)¶	1.09	1.05–1.15
Baseline diabetes mellitus¶	1.61	1.09–2.36
Diabetes mellitus at $k - 1$ #	1.41	0.83–2.40
Aortic calcification at $k - 1$ #	1.32	1.07–1.63

\* The probability of ISH at visit  $k$  is modeled.

† RR, relative risk; CI, confidence interval.

‡ Adjusted for ISH at visit 5 ( $ISH_5$ ) and at visit  $k - 1$  ( $ISH_{k-1}$ ).

§ Adjusted for  $ISH_5$ ,  $ISH_{k-1}$ , and age.

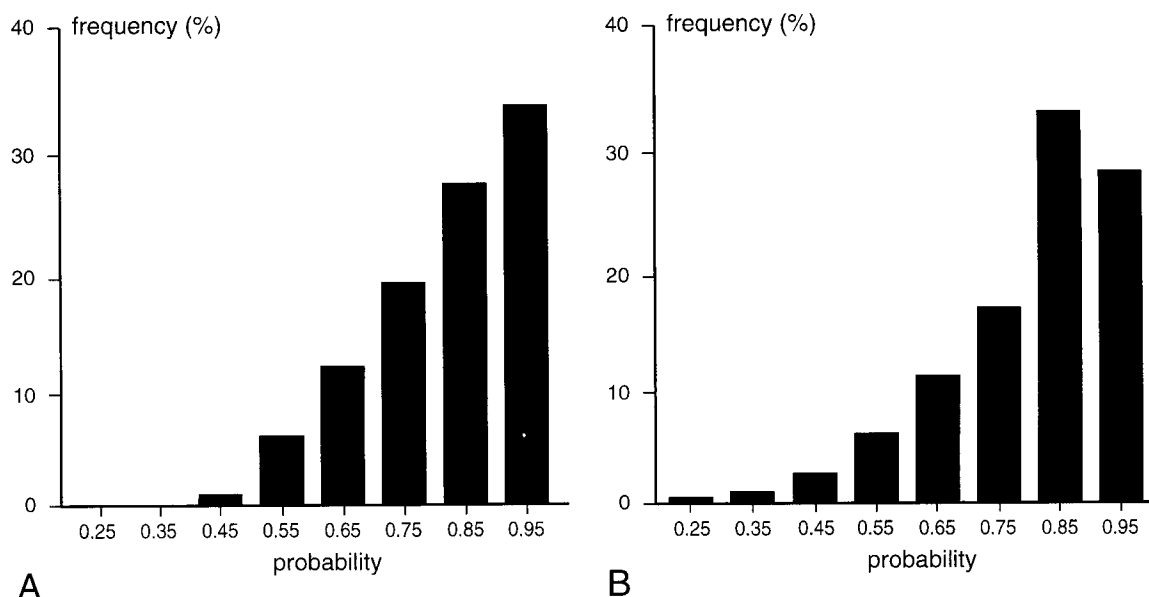
¶ The model (baseline model) includes visit (indicator variables),  $ISH_5$ ,  $ISH_{k-1}$ , age, sex, and the baseline covariates at visit 5 (cigarette smoking, serum cholesterol, metropolitan relative weight, diabetes mellitus, and aortic calcification).

# The model includes all variables of the baseline model and the time-dependent covariate at  $k - 1$ .

number (indicator variables), age, sex, the five baseline covariates and the time-dependent covariates and blood pressure status at  $k$ . Figure 1 gives the probabilities of being free from censoring events through the last visit ( $\hat{K}_i$ ) for subjects who died from cardiovascular disease and subjects censored by end of follow-up. Probabilities could not be computed in 2 percent of subjects because of missing covariate status due to loss to follow-up. These subjects were excluded from the G-analysis.

### G-estimation

Among subjects not censored by competing risks, 113 died from cardiovascular disease during follow-up. For these subjects, we obtained values of  $U_i(\psi)$  by equation 2. We used hypothesized values of  $\psi$  ranging from  $-3.0$  to  $3.0$ . On the basis of  $U_i(\psi)$  and  $C_i$ , we obtained values of  $\Delta_i(\psi)$  by equation 4 for subjects who died from cardiovascular disease or were censored by end of follow-up. We next computed  $\Delta_i(\psi)/\hat{K}_i$  for these subjects and substituted this function for  $U_i(\psi)$  in equation 3. Using generalized estimating equations, we ran three models for the probability of ISH at  $k$  of the form of equation 3. Model 1 included only past history of ISH, entered as  $ISH_5$  and  $ISH_{k-1}$  age, sex, and  $\Delta_i(\psi)/\hat{K}_i$ . Model 2 also included the five baseline covariates measured at visit 5. In model 3, we included the baseline covariates and the time-dependent covariates at  $k - 1$ . Aortic calcification at visits before  $k - 1$  did not predict  $ISH_k$  after adjustment for aortic calcification at  $k - 1$ . Therefore, controlling for aortic calcification at  $k - 1$  will adequately control for time-dependent confounding by aortic calcification. In the presence of time-dependent confounders, model 3 should provide an unbiased estimate under our assumptions. Figure 2 shows how we obtained our point estimate  $\hat{\psi}$  and 95 percent Wald interval for  $\psi_0$ . The point estimate (the value of  $\psi$  where the G-test statistic equals zero) is 0.60, corresponding to a fractional decrease in time to cardiovascular death ( $1 - \exp(-\psi_0)$ ) of 45 percent. The slope ( $\beta$ ) of the G-test statistic evaluated at  $\hat{\psi}$  was estimated by least-square regression of  $Z(\psi)$  on  $\psi$  in that section of the graph that was approximately linear. Specifically, we used values of  $\psi$  from 0.1 to 0.9 with steps of 0.1. The variance of  $\hat{\psi}$ , estimated by squaring the inverse of the slope, is 0.08. The 95 percent confidence interval formed from the Wald statistic,  $\hat{\psi} \pm 1.96 * 1/\beta$ , is 0.03–1.17. The 95 percent test-based confidence interval, as obtained from the graph, is 0.07–1.31. Table 3 gives the point estimates of  $\psi_0$  and 95 percent Wald intervals for the three models. When comparing the results of model 2 with those of model 3, we see that the estimate of  $\psi_0$  does not change when



**FIGURE 1.** Probability of being free from competing risks through the last visit for subjects who died of cardiovascular disease (A) and for subjects censored by end of follow-up (B), Framingham Heart Study, 10-year follow-up using data from 1956 to 1970. Competing risks included development of diastolic hypertension, start of antihypertensive therapy, and death from causes other than cardiovascular disease.

time-dependent covariates are included. Essentially identical intervals were obtained when we used the nonrobust Wald test computed by a standard logistic regression software package for nonclustered data, implying that the effect of within-subject correlation was small in these data.

### The causal rate ratio

We obtained the causal rate ratio from  $\psi_0$  and the shape of the distribution of  $U_i$ . To generate  $U_i$ , we used our estimate of  $\psi_0$ , 0.60. Figure 3 is a Kaplan-Meier plot of  $\log P(U_i > t)$  versus time. Inspection of the graph suggests a Weibull distribution with parameter  $\kappa > 1$ . The point estimate of  $\kappa$  was 1.21 (95 percent confidence interval 1.15–1.28). For this distribution of  $U_i$ , a  $\psi_0$  of 0.60 corresponds to a causal rate ratio of 2.07.

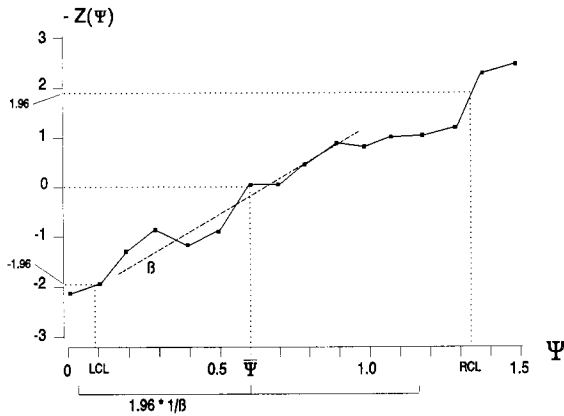
### DISCUSSION

In the present study, we used the method of G-estimation to assess the causal effect of ISH on cardiovascular death. We estimated that ISH decreased the time to cardiovascular death by 45 percent (95 percent confidence interval 3–69). Control for time-dependent covariates did not change the risk estimate. Apparently, the magnitude of time-dependent confounding in our study was small. The presence of calcified plaques in the aorta may be a poor marker of arterial rigidity. New methods for the noninvasive measurement of arterial stiffness have recently been

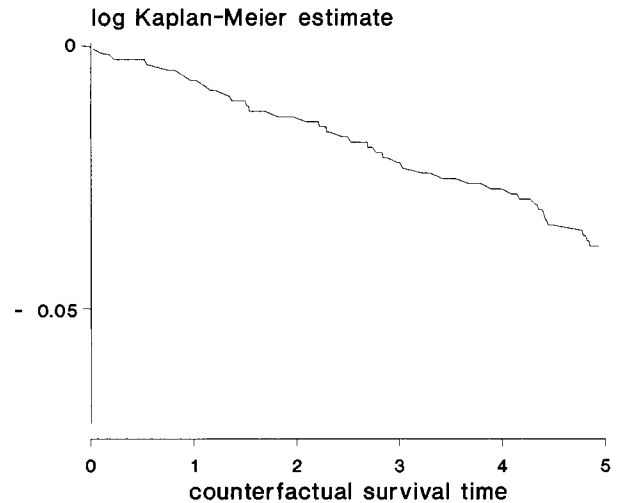
developed (27). However, measurements over time are not yet available in epidemiologic studies. A stronger association was found for diabetes mellitus with cardiovascular death, but the association of the time-dependent covariate with ISH was weaker and nonsignificant. We could have evaluated the magnitude of confounding by standard methods first. However, the best way to examine the consequence of ignoring confounders is to use a method of analysis, such as the G-estimation procedure, that is valid when adjusting for them.

The inference from our analyses is dependent on several assumptions. The first assumption is that the failure time model correctly links an individual's observed failure time and observed ISH history with his unobserved counterfactual failure time,  $U_i$ . In this study, we specified only a simple model. If we do not consider this to be appropriate, we could use a different specification of the model. For example, we could have included  $ISH_{k-1}$  instead of current ISH if we think that ISH status in the interval before current ISH affects cardiovascular death, or we could have included cumulative exposure. Model specification is discussed in more detail by Robins et al. (11, 13, 17).

Second, to test independence of  $ISH_{i,k}$  and  $U_i$ , subjects with and subjects without ISH need to be comparable at  $k$ . This means that we must have available, at each visit  $k$ , data on the history of all time-dependent covariates that are risk factors for death and predict the probability of ISH at  $k$  (the condition of no unmeasured confounders). Whether this is true cannot



**FIGURE 2.** —G-test statistic ( $-Z(\psi)$ ) versus  $\psi$ , Framingham Heart Study, 10-year follow-up using data from 1956 to 1970. For each hypothesized value of  $\psi$ , we ran a model for the probability of isolated systolic hypertension (ISH) at  $k$ , including  $ISH_5$ ,  $ISH_{k-1}$ , age, sex, all baseline covariates at visit 5 (cigarette smoking, serum cholesterol, metropolitan relative weight, diabetes mellitus, and aortic calcification), all time-dependent covariates at  $k - 1$ , and  $\Delta_i(\psi)/K_i$ . See equation 3 where  $U_i(\psi)$  is replaced by  $\Delta_i(\psi)/K_i$ . The point estimate of  $\psi_0$  is the value of  $\psi$  where the G-test statistic equals zero. The slope ( $\beta$ ) of a straight line fit to the graph at  $\psi$  was estimated by regression of  $Z(\psi)$  on  $\psi$  (dotted line), using that section of the graph that was approximately linear. Specifically, we used values of  $\psi$  from 0.1 to 0.9 with steps of 0.1. The 95 percent confidence interval formed from the Wald statistic is  $\psi \pm 1.96 * 1/\beta$ . The left (LCL) and right confidence limits (RCL) of the 95 percent test-based confidence interval are those values of  $\psi$ , where  $-Z(\psi)$  equals  $-1.96$  and  $1.96$ , respectively.



**FIGURE 3.** Kaplan-Meier plot of  $\log P(U_i > t)$  versus time, measured in visit units (2 years), Framingham Heart Study, 10-year follow-up using data from 1956 to 1970. To generate the counterfactual failure times  $U_i$ , we used our estimate of  $\psi_0$ , 0.60 (see equation 1).

be tested by the data. In this study, we assume that this goal has been realized, while recognizing that, in practice, this would never precisely or sometimes even approximately be true.

In the present study, a method to adjust for censoring by competing risks is presented. In our G-estimation procedure, we used weights equal to the inverse of the modeled probability of being free from censoring events through the last visit. We assumed that we had enough information on confounders so

that censoring by competing risks is independent of the time a subject would have died from cardiovascular disease had censoring been prevented. Missing data and loss to follow-up can be treated in a similar way, that is, by specifying a failure time model with failure being the first time for which complete data are unavailable. In the present analysis, missing values at intermediate visits were replaced with values from the closest previous visit. Missing values due to loss to follow-up were present in a small percentage of subjects, and we excluded these subjects from the analysis.

**TABLE 3. Estimates of the causal effect of isolated systolic hypertension (ISH) on cardiovascular death, using G-estimation,\* Framingham Heart Study, 10-year follow-up using data from 1956 to 1970**

	Point estimate of $\psi_0$ †	95% CI†
Model 1‡	0.59	0.03–1.17
Model 2§	0.60	0.03–1.17
Model 3¶	0.60	0.03–1.17

\* The probability of isolated systolic hypertension at visit  $k$  is modeled.

†  $\psi_0$ , unknown parameter; CI, Wald confidence interval.

‡ Model includes visit (indicator variables), ISH at visit 5 ( $ISH_5$ ) and at visit  $k - 1$  ( $ISH_{k-1}$ ), age, sex, and  $\Delta_i(\psi)/K_i$ .

§ Model includes all variables of model 1 and the baseline covariates at visit 5 (cigarette smoking, serum cholesterol, metropolitan relative weight, diabetes mellitus, and aortic calcification).

¶ Model includes all variables of model 2 and all time-dependent covariates at  $k - 1$ .

Some aspects of the G-estimation procedure may be of concern when actually applying the method. Our method of dealing with censoring by end of follow-up artificially censors some subjects with an observed failure time. This did not affect inference in our study, although it may do so in smaller data sets. Nevertheless, the use of artificial censoring is necessary to avoid bias under our assumptions. The method to handle censoring is further complicated in that alternative functions of  $U_i(\psi)$ ,  $ISH_{i,k}$ ,  $L_{i,k}$ , and  $\Delta_i(\psi)$  can be used to replace  $U_i(\psi)$  in equation 3 (15). We used the simplest function with good power,  $\Delta_i(\psi)$ , but we could have used other functions as well. There is an optimal function that minimizes the variance of our G-estimate (11). However, this is a complicated function of the joint distribution of the observed failure times and exposure and covariate histories. As such, estimation of the optimal function would be computationally quite complex. The theory of G-estimation requires data on actual, ungrouped, failure times. Use of grouped failure times, by assuming that any death in

the interval  $(k, k + 1)$  occurred at  $k + 1$ , often results in loss of monotonicity of the graph of  $Z(\psi)$  versus  $\psi$ . Finally, as discussed in Materials and Methods, in the presence of competing risks, it is necessary to use a GEE software package to obtain conservative confidence intervals.

In summary, we estimated the effect of ISH on cardiovascular death under the assumptions specified in this paper. The G-estimation procedure allows for appropriate adjustment of the effect of a time-varying exposure in the presence of time-dependent confounders that are themselves influenced by the exposure. With an expanding body of data on time-dependent exposures and covariates, we think that situations meeting the above conditions will be increasingly encountered in epidemiologic research.

#### ACKNOWLEDGMENTS

This study was supported by NIH/NHLBI grant RO1-HC-40423, contract NO1-HC-38038, the Charles A. Dana Foundation, Merck Sharp and Dohme, and the Framingham Study Visiting Scholar program (ASTRA USA).

The authors thank Dr. Halit Silbershatz for expert help in data management.

#### REFERENCES

- Colandrea MA, Friedman GD, Nichaman MZ, et al. Systolic hypertension in the elderly. An epidemiologic assessment. *Circulation* 1970;XLI:239-45.
- Fineberg MH. Systolic hypertension. Its relationship to atherosclerosis of the aorta and larger arteries. *Am J Med Sci* 1927;173:835-42.
- O'Rourke M. Arterial stiffness, systolic blood pressure, and logical treatment of arterial hypertension. *Hypertension* 1990; 15:339-47.
- Kannel WB, Dawber TR, McGee DL. Perspectives on systolic hypertension. The Framingham Study. *Circulation* 1980;61: 1179-82.
- Kannel WB, Wolf PA, McGee DL, et al. The Framingham Study. Systolic blood pressure, arterial rigidity, and risk of stroke. *JAMA* 1981;245:1225-9.
- Safar ME, London GM. Arterial and venous compliance in sustained essential hypertension. *Hypertension* 1987;10: 133-9.
- Farrar DJ, Bond MG, Sawyer JK, et al. Pulse wave velocity and morphological changes associated with early atherosclerosis progression in the aorta of cynomolgus monkeys. *Cardiovasc Res* 1984;18:107-18.
- Hollander W. Role of hypertension in atherosclerosis and cardiovascular disease. *Am J Cardiol* 1976;38:786-800.
- Kalbfleisch JD, Prentice RL. *The statistical analysis of failure time data*. New York: Wiley, 1980.
- 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. *Math Modeling* 1986;7:1393-512.
- Robins JM, Blevins D, Ritter G, et al. G-estimation of the effect of prophylaxis therapy for *Pneumocystis carinii* pneumonia on the survival of AIDS patients. *Epidemiology* 1992; 3:319-36.
- Robins JM, Blevins M, Ritter G, et al. Errata to "G-estimation of the effect of prophylaxis therapy for *Pneumocystis carinii* pneumonia on the survival of AIDS patients." *Epidemiology* 1993;4:189.
- Robins JM. Estimation of the time-dependent failure time model in the presence of confounding factors. *Biometrika* 1992;79:321-34.
- Robins JM. Analytic methods for HIV treatment and cofactor effects. In: Ostrow DG, Kessler R, eds. *Methodologic issues of AIDS behavioral research*. New York: Plenum, 1993: 213-87.
- 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.
- Robins JM, Greenland S. Adjusting for differential rates of prophylaxis therapy for PCP in high- versus low-dose AZT treatment arms in an AIDS randomized trial. *J Am Stat Assoc* 1994;89:737-49.
- Robins JM. Structural nested failure time models. In: *Survival analysis*. In: Armitage P, Colton T, eds. *Encyclopedia of biostatistics*. Chichester, UK: John Wiley and Sons, 1998.
- Cupples LA, D'Agostino RB. Some risk factors related to the annual incidence of cardiovascular disease and death using pooled repeated biennial measurements. Framingham Heart Study, 30 year follow-up. Bethesda, MD: National Heart, Lung, and Blood Institute, National Institutes of Health, 1987: (Publication no. 87-2703).
- Hyman JB, Epstein FH. A study of the correlation between roentgenographic and post-mortem calcification of the aorta. *Am Heart J* 1954;47:540-3.
- Witteman JCM, Kok FJ, van Saase JLCM, et al. Aortic calcification as a predictor of cardiovascular mortality. *Lancet* 1986;2:1120-2.
- Witteman JCM, Kannel WB, Wolf PA, et al. Aortic calcified plaques and cardiovascular disease (The Framingham Study). *Am J Cardiol* 1990;66:1060-4.
- Cox DR, Oakes D. *Analysis of survival data*. London: Chapman and Hall, 1984.
- Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using a rank preserving structural failure time model. *Comm Stat A* 1991;20:2609-31.
- D'Agostino RB, Lee M-L, Belanger AJ. Relation of pooled logistic regression to time-dependent Cox regression analysis: the Framingham Heart Study. *Stat Med* 1990;9:1501-15.
- Groemping U. *GEE: a SAS macro for longitudinal data analysis*. Dortmund, Germany: Fachbereich Statistik, Universitaet Dortmund, 1993. (Arbeitsbericht no. 13).
- Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol* 1988;128:1185-97.
- Hoeks APG, Brands PJ, Smeets FAM, et al. Assessment of the distensibility of superficial arteries. *Ultrasound Med Biol* 1990;16:121-8.

#### APPENDIX

In this Appendix, we give the SAS code for the G-estimation procedure using a simplified example. We assume we have data collected at 10 visits. The first two visits are used for obtaining data on history of exposure and covariates. Visit 3 is considered to be the baseline visit. Our exposure is denoted by E (ISH in the text). We have only one covariate, denoted by X (e.g., aortic calcification in the text). We have two competing risk events, for example, one fatal (all

causes of death other than the cause of interest) and one nonfatal. To simplify computations, we grouped data on competing risk events such that each event in the interval  $(k, k + 1)$  occurred at  $k + 1$ .

We use a file, called MAIN, containing visits 3–9, with each visit as a separate record. All records of subjects with a competing risk event at or before baseline are deleted. When a competing risk event occurred in the interval  $(k, k + 1)$ , all records after visit  $k$  are deleted. Variables at each record include subject code (ID); visit code (VISIT); exposure at all visits (E3 to E9); current exposure (E); previous exposure, denoted by EP and EPP ( $ISH_{i,k-1}$  and  $ISH_{i,k-2}$  in the text); baseline, current, and previous status of the covariate (X3, X, XP, XPP); the potential censoring time (PC); the time period between visits or, in the case of cardiovascular death, between the last visit and death (PERIOD3 to PERIOD9); survival status (DTH, 0 = surviving through visit 10, 1 = otherwise); competing risk status (CR, 0 = no competing risk event through the last visit, 1 = otherwise); and competing risk event at next visit (CRN, 0 = no event, 1 = fatal event, 2 = nonfatal event).

The first step uses the file MAIN to compute probabilities of being free from competing risk events through the last visit,  $\hat{K}_i$  (referred to as K in code). A polytomous logistic regression with CRN as dependent variable estimates the probabilities of competing risk events at the next visit. The desired probabilities are then readily computed from these fitted probabilities. The probabilities are stored in a file, called CPRNOCR.

Counterfactual failure times are computed using the first record from the file MAIN, which is merged with the file CPRNOCR. All following procedures are written in a SAS macro, called GANAL. In the first section of the code,  $U_i(\psi)$ ,  $C_i(\psi)$ ,  $\Delta_i(\psi)$ , and  $\Delta_i(\psi)/\hat{K}_i$

are computed (referred to as U, C, DELTA, and DELTA\_K in the code). Alternative functions can be specified in this section. The computations are performed for a range of  $\psi$  values (PSI in code), which is specified by the lowest and highest value and increment when invoking macro GANAL (–100, 100, and 10 in code). In the first run, choosing a wide range of PSI values, with large increments, is recommended. The reason for dividing PSI by 0.01 in the computations is purely pragmatic, as only integer values can be used in this macro. The DELTA\_K variables are stored in an array named DKP, created by macro DEFARRAY (which is invoked by macro GANAL), and are written to a file called CFTEMP2. This file is merged with the file MAIN. A GEE logistic regression analysis is then fitted with E as dependent variable and with exposure history, covariate history, and DELTA\_K as independent variables. The link = 3, vari = 3, and corr = 1 options in the SAS code refer to the selection of a logit link function, a binary mean-variance relation, and an independence working covariance matrix, respectively. The fitted model produces a robust test of  $\theta = 0$  ( $\theta$  is the regression coefficient corresponding to DELTA\_K) for each variable DELTA\_K. From the string of G-tests, the G-estimate (referred to as  $\hat{\psi}$  in the text) and the 95 percent confidence interval can be derived as described in the text. We used a GEE SAS macro written by U. Groemping (25); GEE recently became available in SAS PROC GENMOD. When there are no competing risks, a Wald or score test computed by standard logistic regression can be used instead of GEE.

The text of this program and additional SAS codes used to generate the data set from a data set with one record for each ID are available from the first author upon request.

### SAS code for computing probabilities of being free from competing risks ( $\hat{K}_i$ )

```
data TEMP;
  set DATA.MAIN; by ID;
proc logistic descending data = TEMP;
  model CRN = E3 E EP EPP X3 X XP XPP;
  output out = PRTEMP1 pred = PRCR;
```

run;

\* Probabilities are saved in file PRTEMP1 containing two records for each ID. The probability of no competing risk at next visit is  $1 - \text{PRCR}$  for those records in PRTEMP1 where  $\_LEVEL\_ = 1$ ;

```
data PRTEMP2;
  set PRTEMP1;
  if  $\_LEVEL\_ = 1$  then delete;
  PRNOCR =  $1 - \text{PRCR}$ ;
  keep ID VISIT PRNOCR;
```

```

proc sort; by ID VISIT;
run;
* Computing the cumulative probability (K);
data DATA.CPRNOCR;
  set PRTEMP2; by ID;
  LOGPROB = log(PRNOCR);
  if FIRST.ID then SUMLOG = 0;
  if LOGPROB = . then SUMLOG = .;
  else if SUMLOG = . then SUMLOG = .;
  else SUMLOG + LOGPROB;
  K = exp(SUMLOG);
  keep ID K;
  if LAST.ID then output;
run;
* The file CPRNOCR has one record for each ID and includes the variable K;

```

**SAS code for computing  $U_i(\psi)$ ,  $C_i(\psi)$ ,  $\Delta_i(\psi)$ , and  $\Delta_i(\psi)/\hat{K}_i$  and for conducting G-tests of  $\theta = 0$**

```

data FIRST;
  set DATA.MAIN; by ID; if FIRST.ID;
data CFTEMP1;
  merge FIRST DATA.CPRNOCR; by ID;
run;
%macro DEFARRAY(LOW,HIGH,STEP);
  %do I = &LOW %to &HIGH %by &STEP;
    %if &I < 0 %then %let EXI = _%substr(&I,2,%length(&I)-1);
    %else %let EXI = &I;
    DKP&EXI
  %end;
%mend DEFARRAY;
%macro GANAL(LOW,HIGH,STEP);
  data CFTEMP2;
  set CFTEMP1;
  *Computation of DELTA_K for range of PSI values;
  %do PSI = &LOW %to &HIGH %by &STEP;
    if DTH = 1 and CR = 0 then do;
      U = 0;
      %do V = 3 %to 9;
        CT&V = 0;
        if E&V ^ = . then CT&V = PERIOD&V * exp(E&V * (&PSI * 0.01));
        U = U + CT&V;
      %end;
      if &PSI >= 0 then C = PC; else C = PC * exp(&PSI * 0.01);
      if U < C then DELTA = 1;
      if U >= C then DELTA = 0;
      DELTA_K = DELTA/K;
    end;
    else DELTA_K = 0;
    *Putting DELTA_K variables in the array;
    %DEFARRAY(&PSI,&PSI,1) = DELTA_K;
  %end;
  *Storing the array in file CFTEMP2;
  keep ID %DEFARRAY(&LOW,&HIGH,&STEP);
data GANAL;

```

```
merge CFTEMP2 DATA.MAIN; by ID;
%include "GEE.SAS";
*accesses the GEE SAS macro stored in file GEE.SAS;
INTERCEP = 1;
%do PSI = &LOW %to &HIGH %by &STEP;
  %GEE (data = GANAL, missdel = yes, yvar = E, xvar = INTERCEP E3 EP EPP X3 XP XPP
    %DEFARRAY(&PSI,&PSI,1), id = ID, link = 3, vari = 3, corr = 1);
  run;
%end;
%mend GANAL;
%GANAL(-100,100,10);
```