Chapter 14
G-estimation for Accelerated Failure Time Models

Kate Tilling, Jonathan A.C. Sterne, and Vanessa Didelez

14.1 Time-Varying Confounding

There is an increasing interest in life-course epidemiology (Ben-Shlomo 2007; Ben-Shlomo and Kuh 2002), with the quantification of the effects of exposures over long periods of time. For example, several papers recently have examined the effects of socioeconomic position at different stages of life, and changes in that exposure between these stages, on outcomes including risk of stroke and respiratory function, and health behaviours including midlife drinking and smoking patterns (Amuzu et al. 2009; Glymour et al. 2008; Tehranifar et al. 2009; Tennant et al. 2008).

In longitudinal studies, the effects of risk factors on outcome may be estimated in different ways, with different interpretations. The usual approach is to examine the relationship between baseline exposure and rates of disease or death. For reasonably constant exposures, this estimates the cumulative effects of exposure. For example, in a longitudinal study the association between baseline diabetes and subsequent mortality represents the association of lifetime diabetes with mortality. Alternatively we may estimate time-varying effects of exposure. For example, subjects may take up smoking or quit smoking at various stages during the longitudinal study (usually we assume that the exposure level remains constant from one measurement occasion to the next). Here, the time-varying association between smoking and mortality represents the relationship between smoking at a given visit and mortality after that visit. If follow-up is fairly short this represents the instantaneous association between smoking and mortality and can be investigated using standard regression methods (e.g. survival models, structural

K. Tilling (✉) • J.A.C. Sterne
School of Social and Community Medicine, University of Bristol, Bristol, UK
e-mail: kate.tilling@bristol.ac.uk

V. Didelez
School of Mathematics, University of Bristol, Bristol, UK

Y.-K. Tu and D.C. Greenwood (eds.), *Modern Methods for Epidemiology*,
DOI 10.1007/978-94-007-3024-3_14,

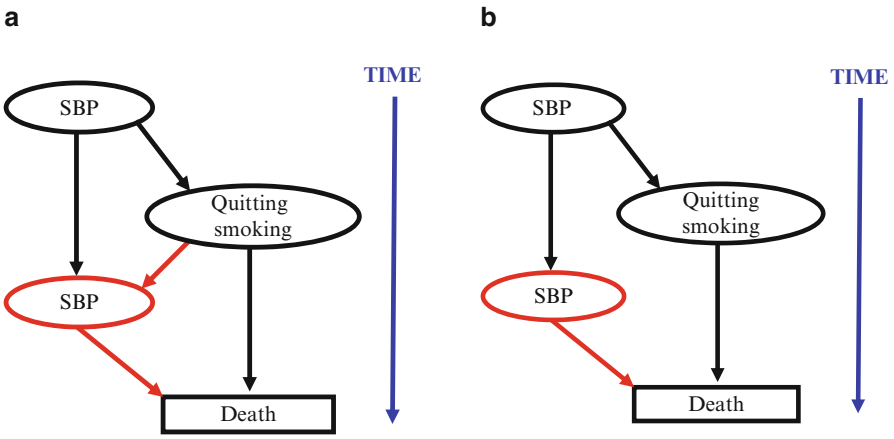

Fig. 14.1 (**a**) Time-varying confounding by SBP of the effect of quitting smoking on mortality. (**b**) Non time-varying confounding by SBP of the effect of quitting smoking on mortality

equation models, etc.). However, increased interest in exposures, confounders and outcomes which vary over time highlights a potential problem, referred to as *time-varying confounding*.

A covariate is a *time-varying confounder* (Mark and Robins 1993; Robins 1986; Young et al. 2010) for the effect of exposure on outcome if it is

1. a time-dependent confounder, i.e. past covariate values predict current exposure and current covariate value independently predicts outcome and also
2. past exposure predicts current covariate value.

As an example, suppose smokers (*exposed*) with high blood pressure are advised to quit smoking, so are less likely to smoke in future (condition 1 above). Suppose also that smoking raises blood pressure (condition 2), and that high blood pressure is a risk factor for death by another pathway other than through smoking (condition 1). In this situation, high blood pressure is a time-varying confounder for the effect of smoking on mortality. Figure 14.1a shows a directed acyclic graph (DAG, see Chaps. 1 and 11) for this example of time-varying confounding. The possible interplay between past and future exposure and confounders makes this very different from the usual definition of a confounder (Chap. 10) where a confounder is always assumed to precede exposure (with DAG for non time-varying confounding shown in Fig. 14.1b). The added complications due to time-varying confounding are twofold. Firstly, if a future covariate is affected by past exposure and independently predicts outcome, then it has the role of a mediator for the effect of past exposure and we do not want to adjust for mediators when estimating the total effect, but at the same time we have to adjust for it because it may confound future exposure and outcome. Secondly, if a covariate is affected by past exposure and other unobserved variables that also predict outcome (e.g. blood pressure may be affected by diet which also predicts survival), then adjusting for this covariate may

introduce selection bias (Hernan et al. 2004), but again, we have to adjust for it if it confounds future exposure and outcome. Hence the question is how to adjust for time-varying confounding without interrupting mediated effects nor introducing selection bias. Standard statistical methods for the analysis of cohort studies (for example Cox or Poisson regression) often get this wrong and yield biased estimates (Robins et al. 1992a), while G-estimation provides a valid method.

We illustrate the problem with an example. When analysing the effect of smoking on mortality we could employ several possible strategies, including: examining the effect of baseline smoking; examining the effect of time-updated smoking; controlling for baseline covariates; and controlling for time-updated covariates.

The unadjusted estimate of the effect of baseline smoking will be biased (favouring smoking, in this case), because those who are both smokers and have high blood pressure (and therefore have the highest mortality risk) will tend to quit subsequently, and thus will reduce their mortality risk. Controlling for baseline covariates such as blood pressure which are measured at the start of the study will still give biased estimates of the effect of smoking, because it ignores the fact that individuals who quit after the start of the study will tend to be those whose blood pressure increased over time.

Controlling for time-updated measurements of covariates such as blood pressure will still give biased estimates of the effect of smoking, because smoking acts on mortality at least partly by raising blood pressure. Controlling for a variable (e.g. blood pressure) which is intermediate on the pathway between the exposure (e.g. smoking) and the outcome (e.g. mortality) will estimate only the direct effect of the exposure (ignoring the effect mediated through the covariate) and may additionally introduce selection bias (Hernan et al. 2004).

Example 1 To illustrate the bias of the usual survival analysis in the situation described above, we simulated data for 2,000 people with four assessment occasions (visits) 3 years apart. Each person had a randomly-generated (log-normally distributed) survival time representing how long they would survive if never exposed, which was then decreased by high blood pressure or smoking. Survival time for a smoker was 0.67 of survival time for a non-smoker with the same covariate history, and survival time decreased by 4% per 1 mmHg increase in current blood pressure. Blood pressure increased by 2 mmHg for current smokers, and by 1 mmHg for ex-smokers (i.e. if an individual smoked at the previous visit but not the current visit, blood pressure was 1 mmHg higher than if they had been a non-smoker at both visits). The odds of smoking were decreased by 0.3 if the participant had high blood pressure at the previous visit. All 2,000 participants were "followed up" until either they died (n = 1,672) or until 3 years after the fourth visit. We took visit 1 to be a baseline visit, and measured time to event/censoring from visit 2. Table 14.1 shows the simulated number at each visit, together with number smoking at that visit and average blood pressure at that visit.

The data were analysed using a Weibull model with the accelerated failure time parameterisation, because this is the parameterisation which corresponds to g-estimation (i.e., calculating the survival ratio rather than the hazard ratio).

Table 14.1 Simulated data for Example 1

Visit	N	N smokers (%)	Mean blood pressure (sd)
1	2,000	140 (7)	142 (4.70)
2	2,000	139 (7)	143 (4.87)
3	1,880	108 (6)	143 (3.80)
4	1,153	67 (6)	141 (4.27)

The accelerated failure time model assumes for the individual failure times T_i with covariates x_i that:

$$T_i = \exp(\theta^T \mathrm{x}_i + \varepsilon_i)$$

where ε_i has a standard extreme value distribution with scale parameter $1/\gamma$, where γ is the shape parameter.

Survival models including current smoking, current smoking and blood pressure, current smoking plus baseline smoking and blood pressure, and smoking and blood pressure at current and previous visits, were all fitted. The model including current smoking only estimated the survival time ratio for smokers compared to non-smokers as 1.14 (95% CI 1.06–1.23), concluding that smoking had little (possibly even a positive) effect on survival. The model including current smoking and current blood pressure estimated the ratio as 0.87 (95% CI 0.84–0.89), that including current smoking and baseline smoking and blood pressure as 0.93 (95% CI 0.90–0.96) and that including all time-updated variables as 0.93 (95% CI 0.91–0.94). Thus all these standard analyses under-estimated the true adverse effect of smoking on mortality (a mortality ratio of 0.67).

14.2 Investigating Time-Varying Confounding

Relationships between time-varying exposures and covariates can be examined using a logistic regression of exposure on concurrent values of the other covariates, values of all exposures and covariates at the previous visit and at baseline (visit 1), and non time-varying covariates. All data from all n visits should be used, so each individual can contribute multiple observations to the model for an exposure. This model will examine condition (1) above. The other part of Condition 1 (whether the covariate affects outcome) can be examined using a model relating outcome to exposure and covariates (e.g. a survival model in the above example, where mortality is the outcome). Condition (2), whether past exposure predicts current covariate values, can be examined using similar logistic regression models of each time-varying covariate on concurrent values of the other covariates and exposure, values of all exposures and covariates at the previous visit and at baseline (visit 1), and non time-varying covariates.

14.2.1 G-estimation

G-estimation of causal effects was proposed by Robins (see e.g. Robins et al. 1992a; Witteman et al. 1998) as one method to allow for confounders which are also on the causal pathway, i.e. time varying confounding. G-estimation has been used in various applications, to estimate the causal association between: quitting smoking and time to death or first CHD (Mark and Robins 1993); isolated systolic hypertension and cardiovascular mortality (Witteman et al. 1998); therapy and survival for HIV-positive men (Joffe et al. 1997, 1998); graft versus host disease and relapse after bone marrow transplants in leukaemia (Keiding et al. 1999); various cardiovascular risk factors and mortality (Tilling et al. 2002); to estimate the total causal effect of highly active antiretroviral therapy (HAART) on the time to AIDS or death among those infected with immunodeficiency virus (HIV) (Hernan et al. 2005); and to correct for non-compliance in clinical trials (Korhonen et al. 1999). G-estimation has also been implemented as a Stata programme (Sterne and Tilling 2002).

14.2.2 Counterfactual Failure Time

The unbiased estimation of causal effects usually requires the assumption of *no unmeasured confounding* (Lok et al. 2004; Robins 1992). Roughly speaking this means that we have measured and included in the model all variables that determine whether a subject is exposed at each measurement occasion and which are also (directly or indirectly) associated with the outcome. G-estimation exploits the assumption of no unmeasured confounding in the following way.

For each subject i, U_i is defined as the time to failure if the subject was not exposed throughout follow-up. This time (the *counterfactual* failure time (Mark and Robins 1993; Robins et al. 1992a; Witteman et al. 1998)) is unobservable for subjects with any exposure. The assumption of no unmeasured confounding (which cannot be tested using the observable data) implies that the exposure for an individual i at a given time will be independent of their counterfactual failure time, U_i, conditional on covariate and exposure history so far. G-estimation proceeds by reconstructing U_i from the observed data and then determining the value of the causal parameter as the one for which this conditional independence is true. An example of this assumption is that, conditional on past weight, smoking status, blood pressure and cholesterol measurements, the decision of an individual to quit or start smoking is independent of what his/her survival time would have been had he/she never smoked. A violation of this assumption would typically occur if the decision depends on unobserved factors, e.g. alcohol consumption, that are informative for the counterfactual survival time U_i. Exposure does not have to be independent of subjects' *current* life expectancy (smokers may choose to quit precisely because they recognise that smoking reduces their life expectancy).

Table 14.2 Simulated data for three individuals in Example 2

Individual	Smoking status				Observed survival time (years)
	Visit 1	Visit 2	Visit 3	Visit 4	
A	1	1			4.88
B	1	1	1	0	11
C	1	1	0	0	14

In its simplest version, G-estimation proceeds by assuming that exposure accelerates failure time by $\exp(-\Psi)$, i.e. $U_i \exp(-\Psi) = T_i^a$ where T_i^a is the survival time for subject i if they are exposed throughout. The actually observed failure time T_i will typically be in between U_i and T_i^a for subjects who have been exposed for some but not all the time. For a given Ψ, the counterfactual survival time $U_{i,\Psi}$ can be calculated backwards from the observed data for subjects who experience an event at time T_i by:

$$U_{i,\Psi} = \int^{T_i} \exp(\Psi \times e_{i,t})\, \mathrm{dt} \tag{14.1}$$

where $e_{i,t}$ is 1 if subject i is actually exposed at time t and 0 if subject i is unexposed. Note that the above model that links the counterfactual survival time $U_{i,\Psi}$ with the observed survival time T_i can be generalised by choosing a more flexible function inside the integral.

For the case where we follow individuals for n follow-up visits (where the first is the baseline visit), we could calculate the counterfactual survival time for subjects who experience an event by:

$$U_{i,\Psi} = \sum_{v=1}^{n} \left[(t_v) \times \exp(\Psi \times e_{i,v})\right]$$

where t_v is the time from visit v to either the event or the next visit.

Example 2 The simulated data on smoking and blood pressure used in example 1 were analysed using g-estimation. We had four visits, of which the first was the baseline. Thus, for a given value of Ψ, the estimated counterfactual survival time for an individual is given by $U_{i,\Psi}$ where

$$U_{i,\Psi} = \sum_{v=1}^{4} \left[(t_v) \times \exp(\Psi \times e_{i,v})\right]$$

where t_v is the time from visit v to either the event or the next visit and $e_{i,v}$ = whether individual i smoked at visit v. Data from three simulated individuals are shown in Table 14.2:

The first individual in this simulated dataset (individual A) was a smoker at visits 1 and 2, and survived for a total of 4.88 years (i.e. 1.88 years from visit 2).

Suppose we assume that smoking halves life expectancy, i.e. $\exp(-\Psi)$=0.5, so Ψ=0.69. Then the counterfactual survival time for this individual at visit 1 (i.e. the length of time they would have survived from visit 1 had they not been a smoker at visit 1 and visit 2) is: $(3 \times \exp(\Psi \times 1)) + (1.88 \times \exp(\Psi \times 1)) = 9.76$ years.

The second individual in this simulated dataset (individual B) was a smoker at visits 1, 2 and 3, then gave up and survived for a further 2 years. Again assuming that smoking halves life expectancy then the counterfactual survival time for this individual at visit 1 (i.e. the length of time they would have survived from visit 1 had they not been a smoker at visits 1, 2 and 3) is: $(3 \times \exp(\Psi \times 1)) + (3 \times \exp(\Psi \times 1)) + (3 \times \exp(\Psi \times 1)) + (2 \times \exp(\Psi \times 0)) = 20$ years.

Thus, for all individuals who are followed up until death, the counterfactual survival time can be calculated in a similar way.

14.2.3 Definition of G-estimation

G-estimation uses the assumption of no unmeasured confounders to estimate the effect of exposure on survival by examining a range of values for Ψ, and choosing the value Ψ_0 for which current exposure is independent of counterfactual survival time U_i (Mark and Robins 1993; Robins 1992; Robins et al. 1992a; Witteman et al. 1998). This can be done by fitting a series of logistic regression models relating current exposure $e_{i,v}$ to $U_{i,\Psi}$, controlling for all confounders (this still assumes that there was no censoring):

$$\text{logit}\ (e_{i,v}) = \mu \mathrm{U}_{\mathrm{i},\Psi} + \sum_{k=0} \alpha_k x_{ik} + \sum_{j} \beta_j c_{ij,v} + \sum_{j=1} \delta_j c_{ij,v-1} + \sum_{j=1} \lambda_j c_{ij,1}\ \mathrm{v} = 2,\ldots, n$$

for different values of Ψ, where $c_{ij,v}$ are the time-varying confounders and x_{ik} the time-invariant confounders. Alternatively, one logistic model could be fitted including data from all visits, with allowance made for clustering within individuals (e.g. by using a GEE). The time-varying confounders may include the values of exposure at previous time-points and at baseline. In fact, the above model for exposure can be generalised and should be chosen according to what is judged appropriate based on subject matter knowledge about the exposure process. For example when exposure is treatment, we may have specific information on the rules according to which treatment was administered. Subjects contribute an observation for each occasion at which their exposure was assessed. The g-estimate Ψ_0 is the value of Ψ for which the Wald statistic of μ in this logistic regression is zero (P value 1, i.e. no association between current exposure and U_{ij,Ψ_0}). The upper and lower limits of the 95% confidence interval for Ψ_0 are the values of Ψ for which the two-sided P-value for the Wald statistic of μ in this logistic regression is 0.05.

This g-estimate Ψ_0 is minus the log of the "causal survival time ratio". Thus $\exp(-\Psi_0)$ estimates the ratio of the survival time of a continuously exposed person to that of an otherwise identical person who was never exposed. If $\exp(-\Psi_0)>1$ then exposure is beneficial (i.e. exposure increases time to the outcome event).

14.3 Censoring – Type I – End of Study

The counterfactual survival time $U_{i,\Psi}$ can only be derived from the observed data for a subject who experiences the event. If the study has a planned end of follow-up (at time C_i for individual i) that occurs before all subjects have experienced the outcome event, not all subjects' counterfactual failure times will be estimable. If C_i is independent of the counterfactual survival time, then this problem can be overcome by replacing $U_{i,\Psi}$ with an indicator variable ($\Delta_{i,\Psi}$) for whether the event would have been observed both if the person had been exposed throughout follow-up and if they had been unexposed throughout follow-up, as described by Witteman et al. (1998).

$$\Delta_{i,\Psi} = \text{ind}(U_{i,\Psi}<C_{i,\Psi})$$

where $C_{i,\Psi} = C_i$ if $\Psi \geq 0$ and $C_{i,\Psi} = C_i \times \exp(\Psi)$ if $\Psi<0$. Thus $\Delta_{i,\Psi}$ is zero for all subjects who do not experience an event during follow-up, and may also be zero for some of those who did experience an event.

Example 3 Continuing with the data from example 1, this study had a planned end of follow-up 12 years after visit 1. Suppose we assume that smoking halves life expectancy, i.e. $\exp(-\Psi)=0.5$, so $\Psi=0.69$. Then for each individual, the indicator variable $\Delta_{i,0.69}$ is equal to 1 if the counterfactual failure time (given $\Psi=0.69$) is less than 12 years and 0 otherwise. The first individual in this simulated dataset (A, above) was a smoker at visits 1 and 2, and survived for 4.88 years from visit 1. The counterfactual survival time for this individual (see example 2) is 9.76 years, and thus the indicator variable $\Delta_{i,0.69}$ takes the value 1 for this individual. The counterfactual failure time for individual B, who smoked at visits 1, 2 and 3 then gave up and survived for another 5 years, was 20 years. Thus the indicator variable $\Delta_{i,0.69}$ takes the value 0 for this individual. Another individual (C) smoked at visits 1 and 2, then gave up and survived until the end of follow-up (dying 14 years after visit 1). As this individual did not experience an event during follow-up, their value for the indicator variable $\Delta_{i,0.69}$ is 0. The value of the indicator variable $\Delta_{i,\Psi}$ can be calculated for all individuals, whether or not they experienced an event during follow-up.

Once the value of the indicator variable has been calculated for each individual, for a given value of Ψ, then the g-estimation can proceed by performing a logistic regression of the exposure of each individual at each timepoint on their

Table 14.3 Simulated data for three individuals in Example 3

Individual	Smoking status				Observed survival time (years)	$\Delta_{i,0.69}$
	Visit 1	Visit 2	Visit 3	Visit 4		
A	1	1			4.88	1
B	1	1	1	0	11	0
C	1	1	0	0	14	0

counterfactual failure time. The data for individuals A, B and C are shown in Table 14.3, assuming that smoking halves life expectancy):

Each individual i then contributes n_i observations to a logistic regression model with exposure as the outcome, where n_i is the number of visits at which that individual has observations. Thus in the example above, individuals A, B and C contribute 1, 3 and 3 observations respectively. In each case, the logistic regression relates their exposure to all their baseline covariates, and previous covariates and exposures, and to the indicator variable for their counterfactual failure time ($\Delta_{i,\Psi}$).

We used g-estimation to estimate the effect of smoking on mortality using the entire simulated dataset. The g-estimate of Ψ was 0.41 (95% CI 0.37–0.44), and the g-estimated survival ratio was 0.66 (95% CI 0.64–0.69) compared to the true value of 0.67. This is closer to the true value than all the other (biased) models (see Example 1), and also has a slightly narrower confidence interval. In this one hypothetical example, g-estimation performs better than the usual survival analysis.

14.4 Censoring – Type II – Competing Risks

Censoring by competing risks can occur when subjects leave the study early or, in the case of cause-specific mortality models, die from other causes. For example, in models where systolic or diastolic blood pressure are the exposures, individuals might be censored when they first reported use of anti-hypertensive medication (Tilling et al. 2002). Subjects could also withdraw from the study because they felt too ill to participate in further follow-ups. In each of these cases, censoring is not independent of the underlying counterfactual survival time. Thus the above method for dealing with censoring by the planned end of a study cannot be used to deal with censoring by competing risks.

As outlined by Witteman et al. (1998), censoring by competing risks is dealt with by modelling the censoring mechanism, and using each individual's estimated probability of being censored to adjust the analysis. This is a similar idea to using weighting for non-response to adjust for missing data (Little and Rubin 2002). Multinomial (if there are several censoring mechanisms) or logistic regression (if there is only one censoring mechanism), based on all available data, is used to relate the probability of being censored at each measurement occasion to the exposure and covariate history. The probability of being uncensored to the end of

the study for each individual is then estimated. The inverse of this probability is used to weight the contributions of individuals to the logistic regression models used in the g-estimation process, to which now only uncensored individuals contribute. This can be done by using probability weights, or by replacing $\Delta_{i,\psi}$ by $\frac{\Delta_{i,\psi}}{p(\text{not censored})}$. This approach means that observations within the same individual are no longer independent, so the logistic regression models for the g-estimation process use robust standard errors allowing for clustering within individuals (using the Huber-White sandwich estimator (Stata Corporation 2007)). This is equivalent to the procedure suggested by Witteman et al., to use a robust Wald test from a generalized estimating equation with an independence working correlation matrix (Witteman et al. 1998). The confidence intervals obtained using this procedure are conservative.

For example, suppose we are examining the effect of systolic and diastolic blood pressure (as exposures) on mortality, and that individuals were censored when they first reported use of anti-hypertensive medication. The probability of being censored at each visit will depend on blood pressure at previous visits, and is likely to be related to other factors also (e.g. smokers may be more likely to have other health problems and therefore to visit the GP). The censoring process is modelled, using logistic regression, with whether the individual was censored (i.e. prescribed anti-hypertensive medication) at each occasion as the outcome. This logistic regression model is then used to derive, for each individual, the probability that they remained uncensored to the end of the study. The inverse of this probability is used to weight all of that individual's contributions to the g-estimation model (using probability weights as before). For example, suppose a person with high initial blood pressure has a chance of 0.25 of being uncensored at the end of the study. In g-estimation the contribution of such a person to the model is multiplied by 4, representing the 'total' of 4 people with high blood pressure, 3 of whom were censored before the end of the study.

14.5 Converting to Survival Analysis

The parameter estimated by the g-estimation procedure, the causal survival time ratio, describes the association between exposure and survival using the accelerated failure time parameterisation. In epidemiology, the more usual parameterisation for survival analysis is that of proportional hazards. It would thus be useful to be able to express the causal survival time ratio in the proportional hazards parameterisation. One obvious way to do this is via the Weibull distribution, as this can be expressed in either parameterisation.

The Weibull hazard function at time t is $h(t) = \phi\gamma t^{\gamma-1}$, where ϕ is referred to as the scale parameter and γ as the shape parameter. If the vector of covariates x_i does

not affect γ, the Weibull regression model can be written as either the usual epidemiological proportional hazards:

$$h(t, \mathrm{x}_i) = h_0(t) \exp(\beta^T \mathrm{x}_i)$$

or accelerated failure time, using the expected failure time:

$$T_i = \exp(\theta^T \mathrm{x}_i + \varepsilon_i)$$

where ε_i has a standard extreme value distribution with scale parameter $1/\gamma$. The Weibull shape parameter γ can thus be used to express results from the accelerated failure time parameterisation as proportional hazards: $\theta = -\beta/\gamma$. If the underlying survival times follow a Weibull distribution, the Weibull shape parameter can be estimated from the survival data and used to express the g-estimated survival ratio as a hazard ratio for the exposure (Witteman et al. 1998).

The 95% confidence intervals for g-estimated effects are generally wider than those for corresponding Weibull estimates, particularly with rare outcomes and for estimates close to 1. This is because G-estimation discards information when censoring, by dichotomising the outcome variable.

Example 4 G-estimation has been used to examine the effects of changes in cardiovascular risk factors in mid-life on all-cause mortality and incidence of coronary heart disease (CHD) (Tilling et al. 2002). Cardiovascular risk factors (systolic and diastolic blood pressure, smoking, diabetes, HDL and LDL cholesterol) were measured four times, with the first measure being used as the baseline in the g-estimation model.

To identify the extent of time-varying confounding, the relationships between each exposure and past and current values of all covariates were examined. This was done using one regression model for each exposure, to which each individual could contribute up to three observations. These models showed that there was substantial time-varying confounding, with inter-relationships among most of the time-varying exposures. Weibull survival analysis (with the accelerated failure time parameterisation) was used to relate all the covariates to survival, and the shape parameter from this model (1.26, 95% CI 1.17–1.36) was later used to express the g-estimated survival ratios as hazard ratios for each exposure.

Separate g-estimation models were fitted for each exposure. In each g-estimation model all risk factors (other than the exposure of interest) were included as time-varying covariates. Baseline variables (e.g. age and sex) were included as non time-varying covariates. In the models for systolic and diastolic blood pressure, individuals were censored when they first reported use of anti-hypertensive medication. The probability of being on anti-hypertensive medication at each visit was dependent on blood pressure at baseline and previous visits, and was also related to baseline and time-varying values of BMI, smoking and diabetes, and to age and sex. This censoring process was modelled, using logistic regression, and the probability of each individual being censored was taken into account in the g-estimation method.

Table 14.4 (modified from (Tilling et al. 2002) with permission of Oxford University Press and the Society for Epidemiologic Research) shows the baseline,

Table 14.4 Baseline and time-varying Weibull survival analysis and G-estimated relations between time-varying cardiovascular risk factors and survival for Atherosclerosis Risk in Communities participants with data from at least the first two visits (1987–1989 and 1990–1993) (Modified with permission of the Oxford University Press and the Society for Epidemiologic Research from Tilling et al. (2002))

Variable	Reference group	Baseline HR	95% CI	Time-varying HR	95% CI	G-estimated HR	95% CI
SBP $\geq$ 140 mmHg	SBP $<$ 140 mmHg	2.08	1.43, 3.03	1.72	1.23, 2.40	1.79	1.38, 2.24
DBP $\geq$ 90 mmHg	DBP $<$ 90 mmHg	1.58	0.79, 3.17	1.91	1.02, 3.56	1.98	0.97, 28.56
Diabetes	No diabetes	2.04	1.67, 2.49	1.26	0.98, 1.62	1.62	1.06, 1.98
BMI (kg/m^2)	BMI 20–30 kg/m^2						
$\leq$20		2.58	1.89, 3.53	3.09	2.03, 4.71	2.07	1.39, 3.64
$\geq$30		1.01	0.85, 1.20	0.83	0.64, 1.07	0.71	0.51, 1.12

HR hazard ratio, *CI* confidence interval, *SBP* systolic blood pressure, *DBP* diastolic blood pressure, *BMI* body mass index, *HDL* high density lipoprotein, *LDL* low density lipoprotein

time-varying and g-estimated hazard ratios for mortality for selected cardiovascular risk factors. The comparisons of the results for the usual survival analysis (relating exposure at baseline to mortality) and g-estimation shed some light on the likely mechanisms for each exposure. Diabetes at baseline was associated with a hazard ratio of 2.04 (Tilling et al. 2002). The g-estimated hazard rate ratio for time-varying diabetes (1.62) was weaker than that for baseline diabetes, indicating that the cumulative effect of diabetes is stronger than the instantaneous effect. The time-varying effect of diabetes was underestimated by the standard analysis (hazard ratio = 1.26). The g-estimated hazard ratio for systolic blood pressure was again weaker than the baseline effect, showing that the effect of blood pressure on mortality was long-term rather than instantaneous. G-estimation and Weibull analysis showed a higher risk of death for those with low BMI and no evidence of increased mortality among subjects with high BMI. The validity of G-estimation depends on there being no unmeasured confounders. Confounders not included here, such as comorbid conditions, may influence the relation between BMI and mortality. Alternatively, BMI may have a cumulative effect, and so short-term changes in weight (assessed by these time-varying models) have a different relation to mortality than long-term weight.

For blood pressure and diabetes, the time-varying effects of exposure were underestimated by the usual survival analysis, whereas the adverse effect of low BMI appeared to be over-estimated by the usual survival analysis. Thus the time-varying confounding present in this example led to biases in the estimation of the effects of time-varying exposures. The confidence intervals for the g-estimated hazard ratios were wider than those for the Weibull estimates, because g-estimation discards information when dichotomising the outcome variable to deal with censoring.

14.6 Extensions to G-estimation

G-estimation (as described above) assumes a binary exposure. The effect of trichotomous exposures on outcome has been estimated using g-estimation and an iterative procedure (Tilling et al. 2002). For each exposure, the middle category was chosen as the reference. One of the other two categories was selected, and the effect of the dichotomous exposure defined by that category and the middle category estimated using g-estimation. This estimate was then included as a fixed value in the g-estimation of the effect of the dichotomous exposure defined by the third category and the middle category. This procedure was iterated to convergence. The standard errors for the effects of variables with three categories estimated in this way may be under-estimated, because each iteration assumes that the effect of the other category on survival is known (rather than estimated). Ideally, both parameters should be estimated simultaneously and a 95% confidence region for their joint distribution calculated. However, this has not yet been carried out in practice. Similarly, there has to date been no extension of g-estimation to continuous exposures.

The parameterisation used in the g-estimation procedure described above assumes that the effect of exposure is both immediate and unlimited. Thus we assume that quitting smoking affects survival from the moment of quitting, and that this effect remains throughout the rest of the non-smoking lifecourse. Alternative models are possible (Lok et al. 2004), for example by generalisations of the integral in Eq. 14.1. They include examining a lagged effect of exposure, or allowing exposure to be related to outcome immediately after exposure, with a lesser effect after a period of time (Joffe et al. 1998). For example, one could hypothesise that the effect of quitting smoking on lung cancer mortality might be lagged, so might not start until 5 years after quitting smoking. The effects of a treatment could also be limited in time – the effect of a particular treatment on outcome may be different in the short and long term (say, before and after 30 months) for example (Joffe et al. 1998). The way in which the counter-factual survival time depends on the exposure can be easily amended to take these alternative hypotheses into account (Joffe et al. 1998).

The use of g-estimation is not restricted to survival outcomes – for example, g-estimation has been used to examine the effects of treatment regimes on non-survival outcomes in randomised clinical trials, allowing for non-compliance (Toh and Hernan 2008). The principle of g-estimation, exploiting the conditional independence between a baseline counterfactual and exposure, has also been used for estimating direct/indirect effects (Goetgeluk et al. 2008), genomic control (Vansteelandt et al. 2009), and for finding optimal treatment strategies (Robins 2004).

14.7 Unmeasured Confounding

G-estimation depends crucially on the assumption of no unmeasured confounding. In particular, it relies on having all variables determining exposure both observed and included in the model. However, in many cohort studies, the factors related to exposure are not measured. For example, when looking at smoking as an exposure there may be many factors related to an individual's decision to quit and success in quitting smoking, which may also be related to the outcome. If these are not all recorded, then there may still be bias in the G-estimation of the effect of smoking. Thus, in order for G-estimation to be used successfully, the factors determining treatment decisions need to be well standardised and well measured. The assumption of no unmeasured confounding is, of course, necessary for the validity of all observational epidemiological analyses.

14.8 Alternatives to G-estimation

Marginal structural models (MSMs) are one type of alternative to g-estimation for analysing longitudinal data (Hernan et al. 2000, 2002; Young et al. 2010). In these models each observation is weighted by the probability of exposure based on past covariate and exposure history, and a model is then fitted to the weighted data and

coefficients interpreted as in a standard analysis. For example, weighted Cox proportional hazards models were used to estimate the joint effect of zidovudine (AZT) and prophylaxis therapy for *Pneumocystis carinii* pneumonia on the survival of HIV-positive men, controlling for time-dependent confounding (Hernan et al. 2001), and the effect of zidovudine therapy on mean CD4 count among HIV-infected men (Hernan et al. 2002; Sterne et al. 2005). The weights were based on the inverse of each patient's probability of the treatment history they actually had, given their covariate history. These inverse probability weights were stabilised and modified to adjust for censoring (Hernan et al. 2001). MSMs were designed to estimate marginal causal parameters and are difficult to adapt to situations where exposure or treatment interacts with covariates. G-estimation in contrast can relatively simply be adapted to include such interactions by modifying the function in the integral and hence the way U_i is calculated back from T_i.

A second alternative to G-estimation is G-computation; also referred to as (parametric) G-formula (Robins et al. 1999; Taubman et al. 2009). The G-formula computes the causal effect of a given exposure or treatment sequence by assuming regression models for all covariates that we wish to adjust for at all measurement time points given the past as well as an outcome regression model, and then integrating out the covariates. In practice this integral needs to be approximated by Monte Carlo simulation. The G-formula is somewhat cumbersome to implement, but has been successfully implemented (Robins et al. 1999; Taubman et al. 2009) and interest in its use is growing (Snowden et al. 2011). The G-formula can also be derived from a decision-theoretic point of view avoiding counterfactuals (Dawid and Didelez 2010).

All three approaches, G-estimation, MSMs, and G-formula, correctly adjust for time-varying confounding but require the same no unmeasured confounding assumption; they differ in that the former two require a valid exposure model in addition to the outcome model, while the latter requires valid models for the time-varying covariates in addition to the outcome model.

14.9 Conclusions and Further Reading

Time-varying confounding may occur in longitudinal studies where exposure and covariates change over time. Where time-varying confounding occurs, it may cause bias in the results of usual survival analyses. G-estimation is one possible method used to overcome this problem, and has been shown to reduce bias in some cases. For those interested in exploring g-estimation further, the following references may be helpful: (Hernan et al. 2005, 2006; Robins 1992, 2008; Robins et al. 1992b, 2007; Tanaka et al. 2008; Yamaguchi and Ohashi 2004; Young et al. 2010). An overview and comparison of three methods of analysing data with time-dependent confounding (marginal structural models and two forms of g-estimation) is provided by Young et al. (2010). A summary of confounding, in particular time-dependent confounding (in the context of marginal structural models) demonstrated using causal diagrams may be found in Robins et al. (2000).

References

Amuzu, A., Carson, C., Watt, H. C., Lawlor, D. A., & Ebrahim, S. (2009). Influence of area and individual lifecourse deprivation on health behaviours: Findings from the British Women's Heart and Health Study. *European Journal of Cardiovascular Prevention and Rehabilitation, 16*(2), 169–173.

Ben-Shlomo, Y. (2007). Rising to the challenges and opportunities of life course epidemiology. *International Journal of Epidemiology, 36*(3), 481–483.

Ben-Shlomo, Y., & Kuh, D. (2002). A life course approach to chronic disease epidemiology: Conceptual models, empirical challenges and interdisciplinary perspectives. *International Journal of Epidemiology, 31*(2), 285–293.

Dawid, A. P., & Didelez, V. (2010). Identifying the consequences of dynamic treatment strategies: A decision theoretic overview. *Statistics Surveys, 4*, 184–231.

Glymour, M. M., Avendano, M., Haas, S., & Berkman, L. F. (2008). Lifecourse social conditions and racial disparities in incidence of first stroke. *Annals of Epidemiology, 18*(12), 904–912.

Goetgeluk, S., Vansteelandt, S., & Goetghebeur, E. (2008). Estimation of controlled direct effects. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 70*, 1049–1066.

Hernan, M. A., Brumback, B., & Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology, 11*(5), 561–570.

Hernan, M. A., Brumback, B., & Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association, 96*(454), 440–448.

Hernan, M. A., Brumback, B. A., & Robins, J. M. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine, 21*(12), 1689–1709.

Hernan, M. A., Hernandez-Diaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology, 15*(5), 615–625.

Hernan, M. A., Cole, S. R., Margolick, J., Cohen, M., & Robins, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety, 14*(7), 477–491.

Hernan, M. A., Lanoy, E., Costagliola, D., & Robins, J. M. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology & Toxicology, 98*(3), 237–242.

Joffe, M. M., Hoover, D. R., Jacobson, L. P., Kingsley, L., Chmiel, J. S., & Visscher, B. R. (1997). Effect of treatment with zidovudine on subsequent incidence of Kaposi's sarcoma. *Clinical Infectious Diseases, 25*(5), 1125–1133.

Joffe, M. K., Hoover, D. R., Jacobson, L. P., Kingsley, L., Chmiel, J. S., Visscher, B. R., & Robins, J. M. (1998). Estimating the effect of zidovudine on Kaposi's sarcoma from observational data using a rank preserving structural failure-time model. *Statistics in Medicine, 17*, 1073–1102.

Keiding, N., Filiberti, M., Esbjerg, S., Robins, J. M., & Jacobsen, N. (1999). The graft versus leukemia effect after bone marrow transplantation: A case study using structural nested failure time models. *Biometrics, 55*(1), 23–28.

Korhonen, P. A., Laird, N. M., & Palmgren, J. (1999). Correcting for non-compliance in randomized trials: An application to the ATBC Study. *Statistics in Medicine, 18*(21), 2879–2897.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken: Wiley.

Lok, J., Gill, R., van der Vaart, A., & Robins, J. (2004). Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. *Statistica Neerlandica, 58*(3), 271–295.

Mark, S. D., & Robins, J. M. (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Statistics in Medicine, 12*(17), 1605–1628.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling, 7*(9–12), 1393–1512.

Robins, J. M. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika, 79*(2), 321–334.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In D. Y. Lin & P. Heagerty (Eds.), *Proceedings of the Second Seattle Symposium on Biostatistics* (pp. 189–326). New York: Springer.

Robins, J. M. (2008). Causal models for estimating the effects of weight gain on mortality. *International Journal of Obesity, 32*(Suppl 3), S15–S41.

Robins, J. M., Blevins, D., Ritter, G., & Wulfsohn, M. (1992a). G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology, 3*(4), 319–336.

Robins, J. M., Mark, S. D., & Newey, W. K. (1992b). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics, 48*(2), 479–495.

Robins, J. M., Greenland, S., & Hu, F. C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association, 94*(447), 687–700.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11*(5), 550–560.

Robins, J. M., Hernan, M. A., & Rotnitzky, A. (2007). Effect modification by time-varying covariates. *American Journal of Epidemiology, 166*(9), 994–1002.

Snowden, J. M., Rose, S., & Mortimer, K. M. (2011). Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology, 173*(7), 731–738.

Stata Corporation. (2007). College Station, Texas.

Sterne, J., & Tilling, K. (2002). G-estimation of causal effects, allowing for time-varying confounding. *The Stata Journal, 2*(2), 164–182.

Sterne, J. A., Hernan, M. A., Ledergerber, B., Tilling, K., Weber, R., Sendi, P., Rickenbach, M., Robins, J. M., & Egger, M. (2005). Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: A prospective cohort study. *The Lancet, 366*(9483), 378–384.

Tanaka, Y., Matsuyama, Y., & Ohashi, Y. (2008). Estimation of treatment effect adjusting for treatment changes using the intensity score method: Application to a large primary prevention study for coronary events (MEGA study). *Statistics in Medicine, 27*(10), 1718–1733.

Taubman, S. L., Robins, J. M., Mittleman, M. A., & Hernan, M. A. (2009). Intervening on risk factors for coronary heart disease: An application of the parametric g-formula. *International Journal of Epidemiology, 38*(6), 1599–1611.

Tehranifar, P., Liao, Y., Ferris, J. S., & Terry, M. B. (2009). Life course socioeconomic conditions, passive tobacco exposures and cigarette smoking in a multiethnic birth cohort of U.S. women. *Cancer Causes & Control, 20*(6), 867–876.

Tennant, P. W., Gibson, G. J., & Pearce, M. S. (2008). Lifecourse predictors of adult respiratory function: Results from the Newcastle Thousand Families Study. *Thorax, 63*(9), 823–830.

Tilling, K., Sterne, J. A., & Szklo, M. (2002). Estimating the effect of cardiovascular risk factors on all-cause mortality and incidence of coronary heart disease using G-estimation: The atherosclerosis risk in communities study. *American Journal of Epidemiology, 155*(8), 710–718.

Toh, S., & Hernan, M. A. (2008). Causal inference from longitudinal studies with baseline randomization. *The International Journal of Biostatistics*, *4*(1), Article 22.

Vansteelandt, S., Goetgeluk, S., Lutz, S., Waldman, I., Lyon, H., Schadt, E. E., Weiss, S. T., & Lange, C. (2009). On the adjustment for covariates in genetic association analysis: A novel, simple principle to infer direct causal effects. *Genetic Epidemiology, 33*(5), 394–405.

Witteman, J. C., D'Agostino, R. B., Stijnen, T., Kannel, W. B., Cobb, J. C., de Ridder, M. A., Hofman, A., & Robins, J. M. (1998). G-estimation of causal effects: Isolated systolic hypertension and cardiovascular death in the Framingham Heart Study. *American Journal of Epidemiology, 148*(4), 390–401.

Yamaguchi, T., & Ohashi, Y. (2004). Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part I: Structural nested models and marginal structural models to test and estimate treatment arm effects. *Statistics in Medicine, 23*(13), 1991–2003.

Young, J. G., Hernan, M. A., Picciotto, S., & Robins, J. M. (2010). Relation between three classes of structural models for the effect of a time-varying exposure on survival. *Lifetime Data Analysis, 16*(1), 71–84.