gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula

Rhian M. Daniel, Bianca L. De Stavola, and Simon N. Cousens
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine
London, UK
Rhian.Daniel@LSHTM.ac.uk

Submitted to *The Stata Journal*

Abstract

This article describes the `gformula` command, an implementation of the g-computation procedure, used to estimate the causal effect of time-varying exposure(s) on an outcome in the presence of time-varying confounders that are themselves also affected by the exposure(s). The procedure can also be used to address the related problem of estimating controlled direct effects and natural direct/indirect effects when the causal effect of the exposure(s) on an outcome is mediated by intermediate variables, and in particular when confounders of the mediator-outcome relationships are themselves affected by the exposure(s). A brief overview of the theory and a description of the command and its options are given, and an illustration using two simulated examples is provided.

Keywords: gformula, causal inference, g-computation formula, time-varying confounding, mediation, direct and indirect effects

1 Introduction

1.1 Time-varying confounding

1.1.1 The setting

Longitudinal studies, where data are collected at several points in time, are common in many areas of research, including epidemiology, clinical trials, ecology, sociology, econometrics, and many more. More specifically, this article deals with the situation in which an explanatory variable (or variables) of interest evolves over time, and is measured at several different fixed points in time on each of a number of units (or subjects). Interest lies in the causal effect of this time-changing explanatory variable on either (1) an outcome of interest, measured at the end of the study, or (2) the time to some event of interest, which could occur at any time during follow-up, but is measured in discrete time, *i.e.* at each visit, when the assessment of whether or not the event has occurred since the last visit is made.

In attempting to measure this causal effect, it is important to consider the role of confounding variables, *i.e.* (informally, with more details below) variables that influence both the explanatory variable and the outcome. Failure to do so typically results in a biased estimator of the causal effect of interest.

Much has been written on the general subject of confounding (Pearl, 2009; Rothman *et al.*, 2008; Morgan and Winship, 2007; Angrist and Pischke, 2009). This paper focusses on the specific problem of time-varying confounders, *i.e.* factors that potentially confound the causal relationship between time-varying explanatory variable and outcome, and that themselves evolve over time and are measured repeatedly throughout the study. In particular, when the time-varying confounder is itself affected by the time-varying explanatory variable of interest, standard methods (*i.e.* regression adjustment) for dealing with confounding can no longer be applied (Robins, 1986; Robins and Hernán, 2009).

This scenario is best illustrated using a causal diagram (Greenland *et al.*, 1999) such as the one depicted in Fig. 1. The arrows in this diagram represent the assumed direction of causal influence. A_0–A_T represent the explanatory variable(s) of interest measured at time-points $0, 1, \ldots, T$. L_0–L_T represent the potential confounder(s) measured at time-points $0, 1, \ldots, T$, where it is assumed that L_t occurs just before A_t. In this diagram, Y is the outcome of interest, measured at the end of the study (visit $T + 1$), and U is a set of unmeasured factors that influence L_0–L_T and Y. Notice that there are no arrows from U to A_0–A_T, and no other common causes (V, say) of A_0–A_T and Y. This represents the (untestable) assumption that, conditional on $\{L_0, L_1, \ldots, L_t\}$ and $\{A_0, A_1, \ldots, A_{t-1}\}$, in the absence of a causal effect of A_t on Y, A_t is independent of Y. This is known as the *no unmeasured confounders* assumption, since it means that at each visit t, a sufficient set of confounders of the relationship between A_t and Y are measured. Notice also

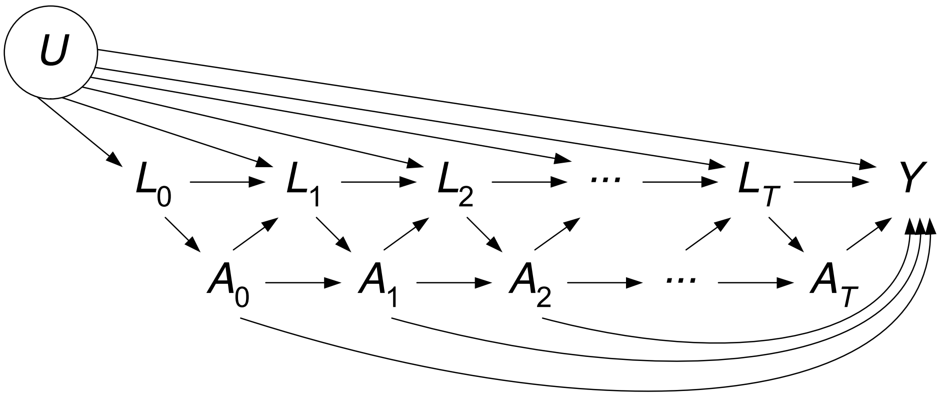

Figure 1: A causal diagram depicting time-varying confounders affected by exposure, when the outcome is measured at the end of follow-up.

the arrows from A_t to L_{t+1}; these represent the fact that the confounder at one time-point may be influenced by the explanatory variable at the previous time-point. (Note that we could have included arrows from A_0 to L_2, from L_0 to A_1 *etc.* These were left out simply to make the diagram more readable, but the omission of the arrows from U to A_0–A_T is crucial).

Suppose instead that the outcome is time-to-event, measured at the discrete time-points 1,2,...,T+1. Then, the appropriate causal diagram is the one shown in Fig. 2, where each Y_t is a binary variable signifying whether or not the event occurred in the time interval $(t-1, t]$. (Again, we could have included arrows from A_0 to L_2, from L_0 to A_1, from L_0 to Y_2 *etc.* but not from U to A_0–A_T).

1.1.2 Limitation of standard methods

The standard method for adjusting for confounding due to L is to condition on L_0–L_T in a regression analysis. Were it not for the arrows from A_t to L_{t+1}, this strategy would succeed in blocking all the so-called *backdoor paths* (Greenland *et al.*, 1999) from A to Y, allowing us to estimate consistently the joint causal effect of A_0–A_T on Y (or $Y_1, \ldots, Y_{T+1}$ in the case of a time-to-event outcome). However, in the situation depicted by Figs. 1 and 2, where the confounder is influenced by past values of the explanatory variable(s), conditioning on L_0–L_T is not valid for two reasons. Consider, for example, the causal effect of A_0 on Y (Fig. 1). By conditioning on L_0, we have successfully blocked the backdoor (non-causal) path $A_0 \leftarrow L_0 \leftarrow U \rightarrow Y$. However, in conditioning on L_1 (and all future L_t) we have blocked the path $A_0 \rightarrow L_1 \rightarrow L_2 \rightarrow \ldots \rightarrow Y$ (and many others) which represents part of the causal effect of A_0 on Y. Furthermore, since U and A_0 both influence L_1, conditioning on L_1 induces a non-causal association between A_0 and U (see Greenland *et al.*, 1999), thereby opening up a new backdoor path from A_0 through U to Y. This is called *collider-stratification bias* and is a form of *selection bias* as classified by Hernán *et al.* (2004). The problem applies similarly to Fig. 2.

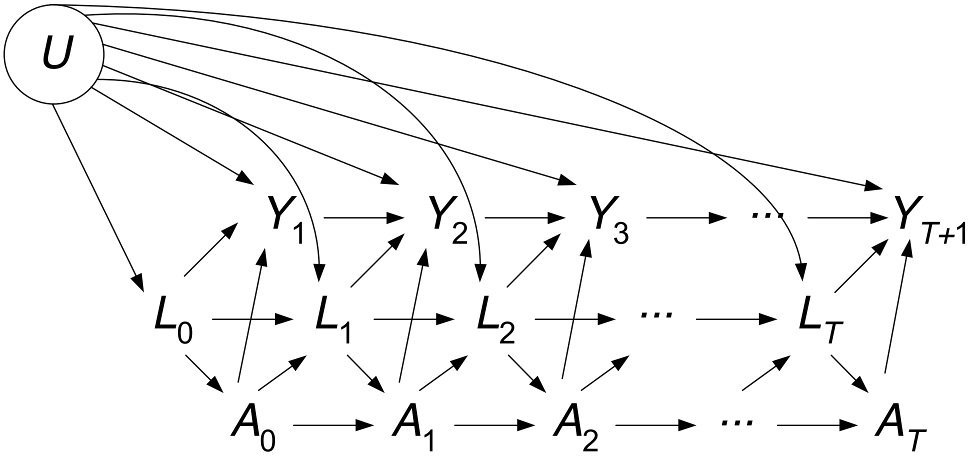

Figure 2: A causal diagram depicting time-varying confounders affected by exposure, when the outcome is time-to-event.

1.1.3 Example I

In a longitudinal study of antiretroviral therapy (ART) in HIV research, A_t is a binary variable indicating whether or not a subject is prescribed ART at time-point t, L_t is CD4 count measured at time t, and Y_t is a binary variable indicating whether or not the subject develops AIDS in the interval $(t-1, t]$. In an observational study, we would expect the decision as to whether or not to treat with ART at a given time-point to be influenced by the current CD4 count of the patient. Also, ART works by raising a patient's CD4 count, and thus adjusting for CD4 in a standard regression analysis is not sensible for the reasons outlined above. We return to this example in section 4.1.

1.2 Mediation

1.2.1 The setting

A substantively different, yet methodologically closely related, problem arises when we wish to decompose the causal effect of an exposure X on an outcome Y into an indirect effect, acting through a mediator M, and a direct effect not mediated by M.

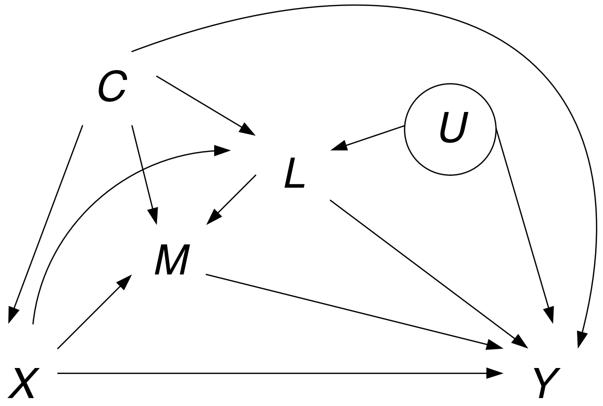

Figure 3: A causal diagram depicting mediation with mediator-outcome confounders affected by the exposure.

1.2.2 Limitation of standard methods

A standard approach in this case would be (i) to fit a regression model for Y conditional on X (and any confounders C of the X–Y relationship) and then (ii) to add M into the same model. Looking at how the coefficient of X changes between models (i) and (ii) is sometimes interpreted as the extent to which the effect of X on Y is mediated by M. More formally, the coefficient of X in model (i) represents the total effect of X on Y, and in model (ii) is often taken to represent the direct effect of X on Y not mediated by M.

This approach is invalid if (as shown in Fig. 3) there are confounders L of the M-Y relationship, since the second model will not consistently estimate the direct effect of X on Y. Conditioning on M induces an association between X and L, opening up a backdoor path from X to L to Y.

Conditioning on L blocks this backdoor path. But if (as shown in Fig. 3) L is affected by X, conditioning on L also blocks the path $X \to L \to Y$, which is part of the direct effect of X on Y (since it is not mediated by M). Thus conditioning on L does not solve the problem arising from conditioning on M whenever L is affected by X.

In addition, the standard regression approach requires that there be no effect modification by X of the effect of M on Y.

1.2.3 Relationship to time-varying confounding

To see the link between the two settings, note that Fig. 3 is the same as Fig. 1, with $T = 1$, $L_0 = C$, $A_0 = X$, $A_1 = M$ and $L_1 = L$. Thus, in the time-varying confounding example, the causal effect of A_0–A_T on Y consists of $T + 1$ direct effects: the direct effect of A_0 on Y not mediated by A_1–A_T, the direct effect of A_1 on Y not mediated by A_2–A_T, and so on.

1.2.4 More on direct/indirect effects

To discuss precisely what we mean by direct and indirect effects, we use some counterfactual notation (Robins and Greenland, 1992; Pearl, 2001). Let $Y(x, m)$ be the potential outcome if, possibly contrary to fact, X were set (by intervention) to x and M were set (by intervention) to m. The *controlled direct effect* (CDE_m) is a comparison of $E\{Y(x, m)\}$ for different values of x, whilst keeping m fixed. For example, if X is univariate and binary, we might specifically consider the controlled direct effect (at m) to be

$$\mathrm{CDE}_m = E\{Y(1, m)\} - E\{Y(0, m)\}$$

Now let $M(x)$ be the potential value of the mediator if, possibly contrary to fact, X were set to x. The *total causal effect* (TCE) is a comparison of $E[Y\{x, M(x)\}]$ for different values of x. Again, for binary X, we would have

$$\mathrm{TCE} = E[Y\{1, M(1)\}] - E[Y\{0, M(0)\}]$$

It would be desirable to use these quantities to infer an indirect effect as the difference between the total effect and the direct effect. The fact that the controlled direct effect is a function of m makes this difficult. CDE_m is potentially different for each value of m.

For this reason, the *natural direct effect* (NDE_{x_0}) is defined to be a comparison of $E[Y\{x, M(x_0)\}]$ for different values of x, keeping x_0 fixed (usually at the 'baseline' value of X, if such a natural choice exists). In other words, it is the effect of X on Y, were M to take on its natural value under the baseline intervention. For binary X, we would have

$$\mathrm{NDE}_0 = E[Y\{1, M(0)\}] - E[Y\{0, M(0)\}]$$

Then the *natural indirect effect* (NIE_{x_1}) can be defined as the difference between the total causal effect and the natural direct effect. Thus it is a comparison of $E[Y\{x_1, M(x)\}]$ for different values of x, whilst keeping x_1 fixed (at a natural choice of 'non-baseline' value). This is best illustrated by thinking again of a binary X, when the natural indirect effect becomes

$$\mathrm{NIE}_1 = E[Y\{1, M(1)\}] - E[Y\{1, M(0)\}]$$

There has been some controversy over whether or not natural direct and indirect effects are well-defined (Robins, 2003; Didelez, 2006; Hafeman, 2009). These (interesting and important) philosophical concerns are beyond the scope of this article, however, and we proceed under the premise that these effects are well-defined in the situations we consider.

1.2.5 Example II

It is widely believed that alcohol consumption has a causal effect on systolic blood pressure (SBP), but the mechanisms through which this causal effect acts are poorly understood. One hypothesis is that alcohol intake affects the level of a liver enzyme, GGT, which, in turn, affects SBP. It would therefore be of interest to know how much of the causal effect of alcohol intake on SBP is mediated by GGT. BMI is thought to affect both GGT and SBP, and socio-economic position (SEP) is thought to affect alcohol intake, BMI and SBP. In addition, alcohol intake has a causal effect on BMI. This situation is depicted by Fig. 3, with X = alcohol intake, M = GGT, Y = systolic blood pressure, L = BMI, and C = SEP. We return to this example in section 4.2.

1.3 A way forward

An alternative to standard regression adjustment is needed to deal with confounding in the two situations described above. One such method is the g-computation procedure, first suggested by Robins (1986) and further discussed by Robins and Hernán (2009) and Taubman *et al.* (2009). In the next section, we give a brief overview of this method, before describing (in section 3) our implementation of it using a new command (`gformula`) in Stata. In section 4, we give an illustration using the two examples described above, before ending (in section 5) with some concluding remarks.

2 The g-computation procedure

2.1 Time-varying confounding

2.1.1 The basic idea

The g-computation procedure works by first modelling the relationships between the variables seen in the observational data. Using these models, we simulate what would have happened to the subjects in the study had the variables A_0–A_T been determined by intervention, rather than been allowed to evolve naturally as in the observational data. The modelling and simulation is carried out 'forward' in time. That is, we start by modelling the time 1 data given the time 0 data, which allows us to simulate the data at time 1 under various hypothetical interventions (on the time 0 exposure) to be compared. Then we model the time 2 data given the time 0 and time 1 data in order to simulate the data at time 2 under the various interventions (on time 0 and time 1 exposures), and so on. All post-baseline confounders and outcome(s) are simulated under each intervention. Causal inference can then be pursued by comparing the outcome(s) under different interventions as if these had been generated from a randomised experiment.

2.1.2 Fitting the models

We specify a parametric model for the conditional distribution of L_1 given L_0 and A_0. (If there are time-fixed confounders, these are included in L_0.) If L_1 is continuous, then $f_{L1|L_0,A_0}\left(l_1\left|l_0,a_0;\alpha_1\right.\right)$ is the probability density function from this model. If L_1 is binary, then $f_{L1|L_0,A_0}\left(l_1\left|l_0,a_0;\alpha_1\right.\right)$ is a conditional probability.

By fitting this model to our observational data on L_1, L_0 and A_0 we obtain estimates $\hat{\alpha}_1$ of α_1.

Similarly, for each $t \in [2,T]$, a model for the conditional distribution of L_t given L_0, A_0, $\ldots$, L_{t-1} and A_{t-1} is specified, and the estimates $\hat{\alpha}_t$ of the parameters α_t from the density/probability $f_{Lt|L_0,A_0,\ldots,L_{t-1},A_{t-1}}\left(l_1\left|l_0,a_0,\ldots,l_{t-1},a_{t-1};\alpha_t\right.\right)$ are obtained from the observational data.

In the case of one outcome Y measured at the end of follow-up, a model for the conditional distribution of Y given L_0, A_0, $\ldots$, L_T and A_T is specified, and the estimates $\hat{\beta}$ of the parameters β from $f_{Y|L_0,A_0,\ldots,L_T,A_T}\left(y\left|l_0,a_0,\ldots,l_T,a_T;\beta\right.\right)$ are obtained from the observational data.

When the outcome is time-to-event, *i.e.* described by a series of binary variables $Y_1, Y_2, \ldots Y_{T+1}$, then, for each $t \in [1, T+1]$, a model for the conditional probability of $Y_t = 1$ given L_0, A_0, $\ldots$, L_{t-1} and A_{t-1} and $Y_{t-1} = 0$ is specified, and the estimates $\hat{\beta}_t$ of the parameters β_t in $f_{Yt|L_0,A_0,\ldots,L_{t-1},A_{t-1}}\left(y_t\left|l_0,a_0,\ldots,l_{t-1},a_{t-1};\beta_t\right.\right)$ are obtained from the subset of the observational data with $Y_{t-1} = 0$ (*i.e.* those still in the risk set).

At present, only `regress` and `logit` are supported by `gformula` in the fitting of these models. There is an option either to fit the models separately at each time-point (although even with this option, the model must be the same at each time-point) or to pool the data across all time-points to estimate the parameters. More details are given in section 3.

2.1.3 Simulating under one hypothetical intervention: the case of a single outcome measured at the end of follow-up

Suppose we wish to simulate what would have happened to the study subjects had the treatment been withheld from all subjects at all times, *i.e.* under the intervention $A_0 = 0$, $A_1 = 0$, $\ldots$, $A_T = 0$.

We use L_0^*, L_1^*, $\ldots$, L_T^* to denote the (simulated) values of L_0, L_1, $\ldots$, L_T under the intervention being considered. L_0 precedes A_0 and is therefore unaffected by any intervention on A_0–A_T. Thus, $L_0^* = L_0$.

L_1^* is simulated from the distribution defined by $f_{L_1|L_0,A_0}\left(l_1 \left| L_0^*, 0; \hat{\alpha}_1\right.\right)$ (see the previous section: **Fitting the models**). In other words, we take the conditional distribution of L_1 given L_0 and A_0 as estimated from the observational data, then we simulate L_1^* from this distribution, after replacing A_0 by 0 and L_0 by L_0^*, *i.e.* the values of A_0 and L_0 under the intervention being considered. If L_1 is continuous, then L_1^* is a stochastic draw from the distribution defined by the density $f_{L_1|L_0,A_0}\left(l_1 \left| L_0^*, 0; \hat{\alpha}_1\right.\right)$. If L_1 is binary, then L_1^* is a stochastic draw from a Bernoulli distribution with probability $f_{L_1|L_0,A_0}\left(1 \left| L_0^*, 0; \hat{\alpha}_1\right.\right)$.

Similarly, L_t^* is simulated from $f_{L_t|L_0,A_0,\ldots,L_{t-1},A_{t-1}}\left(l_1 \left| L_0^*, 0, \ldots, L_{t-1}^*, 0; \hat{\alpha}_t\right.\right)$ for each $t \in [2, T]$.

Finally, Y^* is simulated from $f_{Y|L_0,A_0,\ldots,L_T,A_T}\left(y \left| L_0^*, 0, \ldots, L_T^*, 0; \hat{\beta}\right.\right)$. Y^* is known as a *potential outcome*, since it represents what the outcome would have been under the hypothetical intervention being considered.

Thus, we have simulated all post-baseline variables, including the potential outcome (given the no unmeasured confounders assumption and the modelling assumptions made during the model-fitting stage) under the hypothetical intervention in which treatment is withheld from all subjects at all times. Note that at each stage, the conditional density used for simulation conditions only on past values of the exposure and confounder. This is essentially the difference between this approach and standard regression adjustment for the time-varying confounder. In the latter approach, we condition on the future as well as the past, and this introduces the problems discussed in the **Introduction**.

Since U is unmeasured, the simulation is done marginally over the unobserved distribution of U, but since U is not a confounder of the A–Y relationships, this does not introduce bias (Daniel *et al.*, 2010).

2.1.4 Simulating a time-to-event outcome

In the case of a time-to-event outcome, Y_1^* is simulated from a Bernoulli distribution with probability $f_{Y_1|L_0,A_0}\left(1\,\middle|\,L_0^*, 0; \hat{\beta}_1\right)$. For those with $Y_1^* = 0$, Y_2^* is simulated from a Bernoulli distribution with probability $f_{Y_2|L_0,A_0,L_1,A_1}\left(1\,\middle|\,L_0^*, 0, L_1^*, 0; \hat{\beta}_2\right)$ *etc.* Finally, for those with $Y_T^* = 0$, Y_{T+1}^* is simulated from a Bernoulli distribution with probability $f_{Y_{T+1}|L_0,A_0,\ldots,L_T,A_T}\left(1\,\middle|\,L_0^*, 0, \ldots, L_T^*, 0; \hat{\beta}_{T+1}\right)$. Together, $\left\{Y_1^*, \ldots, Y_{T+1}^*\right\}$ represent the potential time-to-event outcome under the hypothetical intervention which withholds treatment from all subjects at all times.

2.1.5 Comparing many hypothetical interventions

We can change the intervention being studied above from 'never treat' to 'always treat' and repeat the simulation. In this case, we would replace each of A_0–A_T by 1, rather than 0. Similarly, many more hypothetical interventions can be compared using this principle; for example, we could simulate under the intervention 'treat at alternate time-points' or 'treat after time-point 3' *etc.* When A_0–A_T are continuous (and/or multivariate), different hypothetical interventions that set A_0 to a_0, …, A_T to a_T can be compared, for different combinations of values of a_0–a_T.

In the case of a single outcome measured at the end of follow-up we can then compare the hypothetical interventions by calculating the average potential outcome across all subjects for each intervention. Since the average is taken over all subjects, it is marginal over all background variables. In this sense, the g-computation formula should be seen as a form of *standardisation* that is valid for time-varying exposures.

In the case of a time-to-event outcome, the average incidence rate and the cumulative incidence under different hypothetical interventions can be compared, and Kaplan-Meier curves plotted for the different interventions. Again, since these are based on comparing all subjects under different interventions, they are marginal with respect to all other variables.

Under the no unmeasured confounders assumption and the parametric modelling assumptions used in the model-fitting stage, any difference (beyond that expected by finite sample and Monte Carlo simulation error) between the mean potential outcomes / average incidence rates / cumulative incidences / Kaplan-Meier curves can be attributed to the causal effect of the exposure.

2.1.6 The number of subjects simulated

In order to speed up the computation, the simulated subjects can be a random subset of the original dataset. As long as the chosen subset is truly random, this does not introduce bias, but increases the Monte Carlo simulation error and hence decreases precision. This is not to be recommended unless the original dataset is very large, causing the computation time to be unacceptably large.

2.1.7 Standard errors and inference using the bootstrap

Standard errors and confidence intervals are obtained by bootstrapping. The bootstrap samples are taken at the subject level from the original dataset. If the number of Monte Carlo simulations is chosen to be less than the original sample size, the Monte Carlo subset is chosen from the boostrap sample.

2.1.8 Comparing dynamic regimes

The interventions considered so far are all termed *static*, since (in the hypothetical universe in which these interventions are implemented) the treatment trajectory is known fully from the beginning of the (hypothetical) study. Although the hypothetical behaviour of the study participants under these regimes has been constructed using the observational data (in which exposure depends on the past values of the confounder), our aim has been to simulate data free from this dependence.

Another category of interventions is the so-called *dynamic* regimes, where the treatment trajectory is allowed to depend on the confounder trajectory in a pre-specified manner. An example of a dynamic regime in the HIV study would be 'treat once CD4 count falls below x'. The g-computation procedure can be used exactly as outlined above to simulate what would happen to the study participants under different values of x. By trying a range of values of x, an *optimal* regime (in this class) can be sought (*e.g.* in the HIV study, the optimal regime might be defined as the regime which maximises expected AIDS-free survival). Instead of being fixed from the outset, the intervention values A_t^* of A_t now depend on L_0^*, A_0^*, $\ldots$, A_{t-1}^*, L_t^*. Suppose that for a particular subject, $L_0^* > x$,$\ldots$, $L_{t-1}^* > x$, but $L_t^* < x$, then $A_0^* = A_1^* = \cdots = A_{t-1}^* = 0$ and A_t^* is set to 1. This is an example of a *deterministic* dynamic regime, since given past values of the confounder and treatment, the rule defining the dynamic regime assigns a value to the exposure with probability 1. See below (the section on **Simulating under the observational regime**) for an example of a *stochastic* dynamic regime.

2.1.9 Estimating the parameters of a Marginal Structural Model

Thus far, we have described the g-computation procedure as a method for simulating the distribution of Y (or, in the case of a time-to-event outcome, $Y_1, \ldots, Y_{T+1}$) under different hypothetical interventions; and, indeed, this is essentially what it is. However, we may need a more parsimonious way of summarising the comparison. This can be done using a *Marginal Structural Model* (Robins *et al.*, 2000).

A marginal structural model (MSM) expresses some feature of the distribution of a potential outcome as a function of the hypothetical intervention variables. For example, in the case of an outcome measured at the end of follow-up, if $Y\left(a_0, a_1, \ldots, a_T\right)$ is used to denote the potential outcome under the hypothetical intervention that sets the binary variable A_0 to a_0, A_1 to a_1, and so on, then a possible MSM might be expressed as

$$E\left\{Y\left(a_0, a_1, \ldots, a_T\right)\right\} = \gamma + \phi \sum_{t=0}^{T} a_t$$

Thus, given the assumptions made by this MSM, the causal effect of $A_0, \ldots, A_T$ on Y can be summarised in terms of one parameter (ϕ) rather than the (large) set of pairwise comparisons between all the different potential outcomes. By simulating under a large number of different hypothetical interventions, and fitting a regression of Y^* on $\sum_{t=0}^{T} A_t^*$ to the combined simulated dataset (formed by concatenating each of the simulated intervention datasets), estimates of γ and ϕ can be obtained. At present, only MSMs fitted using `regress` and `logit` are supported by `gformula`.

In the case of a time-to-event outcome, if the parameters of the model for Y_1–Y_{T+1} are estimated from a pooled logistic regression over all time-points, then the natural choice of MSM is a marginal structural Cox model (D'Agostino *et al.*, 1990), and `gformula` supports the fitting of such a MSM using `stcox`.

Standard errors and confidence intervals are again obtained by bootstrapping.

According to our definition of a MSM, it is necessarily used to compare static regimes, and thus this option cannot be specified in `gformula` when dynamic regimes are being compared. Dynamic MSMs have been developed (Cain *et al.*, 2010), but are not supported by `gformula` at present.

2.1.10 Dealing with loss to follow-up

In longitudinal studies such as the ones considered here, it is always likely that some subjects drop out before the end of follow-up. Under the assumption that this drop-out occurs at random (Little and Rubin, 2002), *i.e.* that drop-out is conditionally independent of the unobserved data given the observed data (observed prior to drop-out), then such loss to follow-up can be very easily allowed for in the g-computation analysis. Dropping out can be seen as one of the potential treatment trajectories, and then the simulations are made for trajectories such as 'always receive treatment and do not drop out'. The missing at random assumption is then implicit in the no unmeasured confounders assumption. The fact that drop-out need not be explicitly modelled is an example of *ignorability* as defined for likelihood analyses under MAR (Little and Rubin, 2002).

2.1.11 Dealing with censoring due to death

An exception to what is written above occurs when censoring is due to death. It seems unnatural (and indeed potentially misleading) to simulate data under a hypothetical intervention for a subject after the time at which that subject would have died under that intervention. Therefore, survival can be seen as an additional outcome process to be simulated in the same way as described for AIDS-free survival above. First, the question is asked 'did this subject survive the interval $(t-1, t]$?' and then, conditional on the answer to this being simulated as 'yes', the second question 'did the subject develop AIDS in the interval $(t-1, t]$?' is asked, and the answer simulated.

2.1.12 Dealing with missing values in time-fixed variables and intermittent missingness of outcome and time-varying variables using single stochastic imputation

As well as drop-out (where subjects leave the study and never return), longitudinal studies often suffer from intermittent (or *non-monotone*) missingness, that is, some subjects miss a particular visit or visits but return at a subsequent visit. In addition, a subject may have a missing value for a subset of the observations at time-point t, whilst others are observed. Or, a subject may be missing some baseline (time-fixed) variables.

Under the missing at random assumption (which is somewhat contentious for non-monotone patterns of missingness (Robins and Gill, 1997)), such non-monotone patterns of missingness can be dealt with via the method of *multiple imputation using chained equations* (van Buuren *et al.*, 1999). The method as described by van Buuren et al. draws multiple *proper* imputations in the sense described by Little and Rubin (2002). The imputations are termed *proper* since they are not drawn from the distribution of the missing data given the observed under the maximum likelihood estimates of the parameters of this distribution (which would be termed *improper*) but rather from this distribution under Bayesian draws from the posterior distributions of these parameters, with draws from these posterior distributions taken separately for each of the multiple imputations. By drawing multiple proper imputations, Rubin's rules (Little and Rubin, 2002) can then be used to estimate the standard errors of the final parameter estimates.

Since we are not estimating standard errors analytically, but via the bootstrap, the imputation method included in `gformula` is a single stochastic imputation using chained equations. The method is identical to that described by van Buuren *et al.* (1999) except that we draw only one imputation for each missing values, and that imputation is improper. This has been shown to be a valid approach when Rubin's variance estimator is not being used (Tsiatis, 2006, chapter 14).

At present, only `regress`, `logit` and `mlogit` are supported as imputation commands in `gformula`.

2.1.13 Simulating under the observational regime

In addition to simulating what would have happened to the study participants under a number of hypothetical interventions, it is also possible to simulate what would have happened under *no* intervention, that is if the subjects chose their treatments under the same mechanism as was present in the observational data. For this to be possible, a parametric model for treatment assignment given treatment and confounder history must also be specified, and the values of A_1^*–A_T^* under this regime are simulated analogously to what was described for L_1^*–L_T^* above.

In the absence of significant loss to follow-up, a comparison of the actual observed data and the simulated data under the observational regime can give some indication as to the success of the procedure. If these two were very different, it would indicate that at least some of the parametric modelling assumptions, or the no unmeasured confounding assumption, do not hold. But good agreement between the actual observed data and the simulated data under the observational regime does not guarantee that the assumptions hold.

The observational regime is an example of a dynamic regime (see above). In fact, since each A_t^* is a draw from a distribution (to mimic the variation seen in an observational setting), this is an example of a *stochastic* dynamic regime.

A comparison between a given intervention and the observational regime is often of interest when assessing the likely impact of such an intervention if implemented in the population being studied (Taubman *et al.*, 2009).

2.2 Mediation

The g-computation procedure for the mediation example works in a similar way, except that for the natural direct and indirect effects, simulations under different hypothetical interventions need to be combined. Suppose that X is binary, then M is simulated under both $X = 1$ and $X = 0$, giving $M(1)$ and $M(0)$, respectively. To simulate $Y\{1, M(0)\}$ (needed to estimate the natural direct effect), X is set to 1 at the same time as M is set to the simulated value under the intervention $X = 0$, *i.e.* $M(0)$.

If X is not binary and/or if X is multivariate, there may not be a natural comparison (such as 1 vs. 0) for calculating the total causal effect, controlled direct effect or natural direct/indirect effects. In this case, the formulae in section 1.2 are replaced by

$$\text{CDE}_m = E\{Y(X,m)\} - E\{Y(0,m)\}$$

$$\text{TCE} = E\{Y(X, M(X))\} - E\{Y(0, M(0))\}$$

$$\text{NDE}_0 = E[Y\{X, M(0)\}] - E[Y\{0, M(0)\}]$$

and

$$\text{NIE}_X = E[Y\{X, M(X)\}] - E[Y\{X, M(0)\}]$$

where 0 is still the 'baseline' value(s) of X, but is now compared with the distribution of X arising naturally in the observational data. Such a comparison often corresponds to the causal question of interest.

Missing data in any of the variables can be dealt with via single stochastic imputation using chained equations as described above.

3 The gformula command

3.1 Syntax

```
gformula mainvarlist [if] [in], outcome(varname) commands(string) equations(string)
[idvar(varname) tvar(varname) varyingcovariates(varlist) intvars(varlist) interventions(string)
dynamic eofu pooled death(varname) derived(varlist) derrules(string) fixedcovariates(varlist)
laggedvars(varlist) lagrules(string) msm(string) mediation exposure(varlist) mediator(varlist)
control(string) baseline(string) base_confs(varlist) post_confs(varlist) impute(varlist)
imp_cmd(string) imp_eq(string) imp_cycles(#) simulations(#) samples(#) seed(#) obe all graph
saving(string) replace]
```

where *mainvarlist* contains all the variables to be used by the command. Please note that neither the abbreviation of variable names nor the use of variable lists (such as `x1-x3` to denote `x1 x2 x3`) is supported. Categorical variables should be listed using only their names (*e.g.* `agecat`) without using the prefix `i.` (*e.g.* `i.agecat`).

3.2 The data structure

For the time-varying confounding option (as opposed to the mediation option—see below), the data must be in long format (see [D] **reshape**), *i.e.* there should be a separate record for each subject at each time-point. If the outcome is time-to-event, the outcome data for each subject should be given as a series of binary variables measured at each time-point, as suggested by Fig. 2. No records should be included in the dataset for subjects who have been censored before that time, due to death or loss to follow-up, or (in the case of a time-to-event outcome) due to having experienced the event before that time.

Any value which is to be imputed (including those at intermittent missing visits, for which a record must be included) should be denoted by a “`.`”, according to Stata’s convention.

For the mediation option, there should be exactly one record per subject. Again, missing values to be imputed should be denoted by a “`.`”.

Examples of how the data should be structured in each situation are given in section 4.

3.3 Options

3.3.1 Time-varying confounding options

`outcome(`*varname*`)` specifies that *varname* is the outcome variable.

`commands(`*string*`)` specifies which command (either `regress` or `logit`) should be used when fitting each of the parametric models. The variable name is followed by a colon (`:`), which is followed by the command name, with a comma (`,`) separating the different variables (see the example syntax in section 4).

Commands should be specified for the models for the outcome variable, time-varying confounders and the time-varying exposure. If there is censoring due to death, then the command used for the model for death should also be specified.

For a time-to-event outcome (and death, if applicable), `logit` is the sensible command to choose, since the outcome (or death) is given (and simulated) as a sequence of binary variables.

`equations(`*string*`)` specifies the right-hand side of the equations used when fitting the models listed above. The name of the dependent variable is followed by a colon (`:`), which is followed by the list of independent variables. A comma (`,`) should separate the equations for the different dependent variables (see the example syntax in section 4).

Since the data are stored in long format, *lagged* variables will need to be used (see below) to incorporate the dependence on data from previous visits.

The equation for any particular variable, for example a time-varying confounder L, must be the same at each visit.

Variables that are to be treated as categorical variables on the RHS of any equation should be preceded by "`i.`".

`idvar(`*varname*`)` specifies that *varname* is the numeric variable identifying the subject.

`tvar(`*varname*`)` specifies that *varname* is the numeric variable identifying the time-point.

`varyingcovariates(`*varlist*`)` specifies that *varlist* are the time-varying covariates. If lagged versions of these variables are to be used, only the unlagged versions should be included in this list.

`intvars(`*varlist*`)` specifies that *varlist* are the variables on which interventions are to be specified. If lagged versions of these variables are to be used, only the unlagged versions should be included in this list.

`interventions(`*string*`)` specifies the exact interventions to be compared. Different interventions should be separated by a comma (`,`) and different commands within one intervention should be separated by a backwards slash (\) (see the example syntax in section 4).

`dynamic` specifies that the regimes to be compared are dynamic. If this option is not specified, it is assumed that the regimes to be compared (except for the observational regime) are all static.

`eofu` specifies that the outcome is measured only at the end of follow-up. If this option is not specified, it is assumed that the outcome is time-to-event.

`pooled` specifies that the models defined by the `command` and `equation` options above (along with the models defined by the `imp_cmd` and `imp_eq` options below, if applicable) should be fitted to data from all visits at once, pooling across time-points. If this option is not specified, the models are fitted separately at each visit.

death(*varname*) gives the name of the variable (a sequence of binary variables at each time-point) which takes the value 0 if a subject is still alive at that time-point and 1 if a subject died between the previous and current time-points. (No further records following death should be included in the original dataset.)

It is assumed that all censoring (before the final visit) where death=0 is due to loss to follow-up. Simulations are then drawn to mimic a situation in which there are deaths but no losses to follow-up.

If the death option is not specified, all censoring (before the final visit) is assumed to be due to loss to follow-up and simulations are drawn to mimic no losses to follow-up.

derived(*varlist*) lists all the variables which are to be derived from other variables, such as interactions. Lagged variables themselves should not be included here, but variables derived using one or more lagged variables should be included. The derived variables must exist in the original dataset.

derrules(*string*) describes how the derived variables are to be obtained from the other variables. For example, if the variable al is to be created as the product of a and l, the code is derrules(al:a*l) (and al should be included in derived(*varlist*) above). The rules for generating more than one derived variable should be separated using a comma.

fixedcovariates(*varlist*) lists the time-fixed covariates. These do not depend on the time-varying exposure and thus are not simulated.

laggedvars(*varlist*) lists the lagged variables. The lagged variables must exist in the original dataset.

lagrules(*string*) gives further details of the lagged variables. For example, if the variable a_lag is the lagged version of a, and a_lag2 is the double-lagged version of a, this would be denoted as lagrules(a_lag:a 1,a_lag2:a 2).

msm(*string*) specifies the form of the marginal structural model, for example, msm(regress y a_lag a_lag2) or msm(stcox a_lag a_lag2). Only regress, logit and stcox are supported at present. This option cannot be specified in conjunction with dynamic.

impute(*varlist*) gives a list of the variables that contain missing values to be imputed via the method of single stochastic imputation using chained equations.

`imp_cmd(`*string*`)` specifies which command (either `regress`, `logit` or `mlogit`) should be used when fitting each of the imputation models. The syntax is the same as for the `commands` option described above.

`imp_eq(`*string*`)` specifies the RHS of each of the equations to be used for fitting each of the imputation models. The syntax is the same as for the `equations` option described above.

`imp_cycles(`#`)` specifies the number of cycles of chained equations to be used in the imputation procedure. The default is 10.

`simulations(`#`)` specifies the size of the Monte Carlo simulated dataset. The default is the same size as the observed dataset, but for computational reasons, it can be smaller.

`samples(`#`)` specifies the number of bootstrap samples. The default is 1,000.

`seed(`#`)` sets the random-number seed to #.

`all` specifies that all bootstrap confidence intervals are to be displayed (normal, percentile, bias corrected, and bias corrected and accelerated). The default is to give normal-based bootstrap confidence intervals only. See [R] **bootstrap**.

`graph` specifies that a Kaplan-Meier plot of the survival curves under each intervention be displayed. This option is only relevant for a time-varying confounding analysis with a time-to-event outcome.

`saving(`*string*`)` saves the dataset containing the original observational data and all the Monte Carlo simulations in a Stata dataset named *string*. The dataset contains a variable `_int` which takes the value 0 for the observational data, the value 1 for the simulations corresponding to intervention 1, and so on for each of the m specified interventions. Finally, the Monte Carlo simulations under the 'no intervention' regime appear at the end of the dataset, with `_int` taking the value $m + 1$.

`replace` specifies that if the .dta file given in the option `saving(`*string*`)` already exists, it should be overwritten.

3.3.2 Mediation options

`mediation` specifies that the analysis is a mediation analysis. If this option is not specified, then a time-varying confounding analysis is assumed.

`outcome(`*varname*`)` specifies that *varname* is the outcome variable.

`commands(`*string*`)` specifies which command (either `regress` or `logit`) should be used when fitting the parametric models used as a basis for simulation. The variable name is followed by a colon (`:`), which is followed by the command name, with a comma (`,`) separating the different variables (see the example syntax in section 4).

Models must be specified for the mediator(s), outcome, and the post-baseline confounders of the mediator-outcome relationship that are affected by the exposure.

`equations(`*string*`)` specifies the right-hand side of the equations used when fitting the models listed above. The name of the dependent variable is followed by a colon (`:`), which is followed by the list of independent variables. A comma (`,`) should separate the equations for the different dependent variables (see the example syntax in section 4).

Variables that are to be treated as categorical variables on the RHS of any equation should be preceded by "`i.`".

`derived(`*varlist*`)` lists all the variables which are to be derived from other variables, such as interactions.

`derrules(`*string*`)` describes how the derived variables are to be obtained from the other variables. For example, if the variable `al` is to be created as the product of `a` and `l`, the code is `derrules(al:a*l)` (and `al` should be included in `derived(`*varlist*`)` above). The rules for generating more than one derived variable should be separated using a comma.

`exposure(`*varlist*`)` specifies the exposure variable(s).

`mediator(`*varlist*`)` specifies the mediator variable(s).

`control(`*string*`)` specifies the value(s) at which the mediator(s) should be controlled for the controlled direct effect (see the example syntax in section 4). If this option is not specified, only natural direct/indirect effects are estimated.

`baseline(`*string*`)` specifies the value(s) of the exposure(s) to be taken as baseline value(s) (see the example syntax in section 4). `base_confs(`*varlist*`)` specifies the confounder(s) of the exposure-outcome relationship(s).

`post_confs(`*varlist*`)` specifies the confounder(s) of the mediator-outcome relationship(s).

`impute(`*varlist*`)` gives a list of the variables that contain missing values to be imputed via the method of single stochastic imputation using chained equations.

`imp_cmd(`*string*`)` specifies which command (either `regress`, `logit` or `mlogit`) should be used when fitting each of the imputation models. The syntax is the same as for the `commands` option described above.

`imp_eq(`*string*`)` specifies the RHS of each of the equations to be used for fitting each of the imputation models. The syntax is the same as for the `equations` option described above.

`imp_cycles(`#`)` specifies the number of cycles of chained equations to be used in the imputation procedure. The default is 10.

`simulations(`#`)` specifies the size of the Monte Carlo simulated dataset. The default is the same size as the observed dataset.

`samples(`#`)` specifies the number of bootstrap samples. The default is 1,000.

`seed(`#`)` sets the random-number seed to #.

`obe` specifies that there is only one binary exposure, and that the comparisons should be made between $X = 1$ and $X = 0$. If this is not specified, comparisons are made between the natural distribution of X in the observed data, and the baseline value(s).

`all` specifies that all bootstrap confidence intervals are to be displayed (normal, percentile, bias corrected, and bias corrected and accelerated). The default is to give normal-based bootstrap confidence intervals only. See [R] **bootstrap**.

`saving(`*string*`)` saves the dataset containing the original observational data and all the Monte Carlo simulations in a Stata dataset named *string*.

`replace` specifies that if the .dta file given in the option `saving(`*string*`)` already exists, it should be overwritten.

4 Illustration using two simulated examples

4.1 Example I: time-varying confounding

4.1.1 The data

Two datasets are simulated with $T = 9$, according to the description given in section 1.1.3. In the first dataset, A_0–A_9 are binary treatment variables with $A_t = 1$ if a subject is prescribed antiretroviral therapy (ART) at visit t, and 0 otherwise. L_0–L_9 are the values of the logarithm of CD4 count at each visit. Y_1–Y_{10} are binary variables, where $Y_t = 1$ if a subject develops AIDS during the time-interval $(t-1, t]$ and 0 otherwise. All subjects are AIDS-free at baseline (hence Y_1 is the first recorded measurement of Y) and if $Y_t = 1$, no records are included for that individual from time $t + 1$ onwards. Here are the data for the first three subjects in the first dataset.

```
+-------------------------------------------------------------------+
|  id    t   y          l   a   cuma   a_lag   cuma_lag       l_lag |
|-------------------------------------------------------------------|
|   1    0   .   5.195231   1      1       0          0           0 |
|   1    1   0   5.524413   1      2       1          1    5.195231 |
|   1    2   0   5.813174   0      2       1          2    5.524413 |
|   1    3   0   5.322465   0      2       0          2    5.813174 |
|   1    4   0   4.547185   1      3       0          2    5.322465 |
|   1    5   0   4.963298   0      3       1          3    4.547185 |
|   1    6   0   4.389211   0      3       0          3    4.963298 |
|   1    7   0   3.977597   1      4       0          3    4.389211 |
|   1    8   0   4.533944   1      5       1          4    3.977597 |
|   1    9   0   5.023493   1      6       1          5    4.533944 |
|   1   10   0          .   .      .       1          6    5.023493 |
|-------------------------------------------------------------------|
|   2    0   .   4.686166   0      0       0          0           0 |
|   2    1   0    4.05956   0      0       0          0    4.686166 |
|   2    2   0   3.569694   0      0       0          0     4.05956 |
|   2    3   0   3.038292   1      1       0          0    3.569694 |
|   2    4   0   3.584502   1      2       1          1    3.038292 |
|   2    5   0   4.249334   1      3       1          2    3.584502 |
|   2    6   0   4.817859   1      4       1          3    4.249334 |
|   2    7   1   5.353656   1      5       1          4    4.817859 |
|-------------------------------------------------------------------|
|   3    0   .   6.051494   0      0       0          0           0 |
|   3    1   0   5.407419   0      0       0          0    6.051494 |
|   3    2   0   4.841232   0      0       0          0    5.407419 |
|   3    3   0   4.412408   0      0       0          0    4.841232 |
|   3    4   0   4.074246   0      0       0          0    4.412408 |
|   3    5   0   3.754061   1      1       0          0    4.074246 |
|     3     6   0   4.415367   1       2         1           1   3.754061 |
|     3     7   0   4.870831   0       2         1           2   4.415367 |
|     3     8   0   4.272728   0       2         0           2   4.870831 |
|     3     9   0   3.979522   1       3         0           2   4.272728 |
|     3    10   0          .   .       .         1           3   3.979522 |
|-------------------------------------------------------------------------|
```

Subjects 1 and 3 remained AIDS-free until the end of follow-up. Subject 2 developed AIDS between times 6 and 7. `a_lag` is the lagged version of `a`, *i.e.* it contains the previous value of `a`, except at time 0 when `a` is 0 for all subjects. Similarly, `l_lag` is the lagged version of `l`. `cuma` at time t is the sum of all the values of `a` for that subject up to and including time t, and `cuma_lag` is its lag.

The data (consisting of 1,000 subjects) were generated as follows.

- U is a normal random variable with mean 0 and variance 0.25.
- L_0 is a normal random variable with mean $5.5 + U$ and variance 0.04.
- A_0 is generated from a Bernoulli distribution with probability

$$\frac{\exp\left(5 - L_0\right)}{1 + \exp\left(5 - L_0\right)}$$

- Then for each $t \in [1, 9]$, and for those with $Y_{t-1} = 0$, Y_t, L_t and A_t are generated as follows. Y_t is generated from a Bernoulli distribution with probability

$$\frac{\exp\left(-8 + L_{t-1} - 0.3\sum_{s=0}^{t-1} A_s - U\right)}{1 + \exp\left(-8 + L_{t-1} - 0.3\sum_{s=0}^{t-1} A_s - U\right)}$$

  L_t is generated from a normal distribution with mean $0.9L_{t-1} + A_{t-1} + 0.1U$ and variance 0.01. A_t is generated from a Bernoulli distribution with probability

$$\frac{\exp\left(A_{t-1} + 4.5L_t\right)}{1 + \exp\left(A_{t-1} + 4.5L_t\right)}$$

- Finally, for those with $Y_9 = 0$, Y_{10} is generated from a Bernoulli distribution with probability

$$\frac{\exp\left(-8 + L_9 - 0.3\sum_{s=0}^{9} A_s - U\right)}{1 + \exp\left(-8 + L_9 - 0.3\sum_{s=0}^{9} A_s - U\right)}$$

The second dataset is generated in exactly the same way, except that there is censoring, both due to death and due to losses to follow-up. Everyone is observed at time 0. Thereafter, loss to follow-up at time t is generated as a Bernoulli random variable with mean

$$\frac{\exp\left(-6 + 0.5L_{t-1} - 0.1\sum_{s=0}^{t-1} A_s - U\right)}{1 + \exp\left(-6 + 0.5L_{t-1} - 0.1\sum_{s=0}^{t-1} A_s - U\right)}$$

If loss to follow-up has not occurred, then death is simulated at time t as a Bernoulli random variable with mean

$$\frac{\exp\left(-10 + L_{t-1} - 0.3\sum_{s=0}^{t-1} A_s - U\right)}{1 + \exp\left(-10 + L_{t-1} - 0.3\sum_{s=0}^{t-1} A_s - U\right)}$$

If neither death nor loss to follow-up has occurred, then Y_t, L_t and A_t are generated as shown above.

Here are the data for four subjects from this second dataset (`d` is the variable denoting death).

```
+-----------------------------------------------------------------------+
|   id    t   d   y          l   a   cuma   a_lag   cuma_lag      l_lag |
|-----------------------------------------------------------------------|
|    1    0   .   .   5.195231   1      1       0          0          0 |
|    1    1   0   0   5.594172   0      1       1          1   5.195231 |
|    1    2   1   .          .   .      .       0          1   5.594172 |
|-----------------------------------------------------------------------|
|    2    0   .   .   4.686166   0      0       0          0          0 |
|    2    1   0   0   4.151603   1      1       0          0   4.686166 |
|    2    2   0   0   4.702286   1      2       1          1   4.151603 |
|    2    3   0   0   5.079803   1      3       1          2   4.702286 |
|    2    4   0   0   5.504702   0      3       1          3   5.079803 |
|    2    5   0   0   4.979135   1      4       0          3   5.504702 |
|    2    6   0   0     5.3479   1      5       1          4   4.979135 |
|    2    7   0   0   5.840528   0      5       1          5     5.3479 |
|    2    8   0   0    5.09903   0      5       0          5   5.840528 |
|    2    9   0   0   4.505893   0      5       0          5    5.09903 |
|    2   10   0   0          .   .      .       0          5   4.505893 |
|-----------------------------------------------------------------------|

|-----------------------------------------------------------------------|
|    9    0   .   .   5.392478   0      0       0          0          0 |
|    9    1   0   0   4.849835   1      1       0          0   5.392478 |
|    9    2   0   0   5.218192   1      2       1          1   4.849835 |
|    9    3   0   0   5.610555   0      2       1          2   5.218192 |
|    9    4   0   0    5.12606   0      2       0          2   5.610555 |
|    9    5   0   0   4.474302   1      3       0          2    5.12606 |
|    9    6   0   1    5.07163   1      4       1          3   4.474302 |
|-----------------------------------------------------------------------|

|-----------------------------------------------------------------------|
|   30    0   .   .   5.347866   1      1       0          0          0 |
|   30    1   0   0   5.937762   0      1       1          1   5.347866 |
|   30    2   0   0   5.463366   0      1       0          1   5.937762 |
|   30    3   0   0   4.786379   0      1       0          1   5.463366 |
|   30    4   0   0   4.581161   1      2       0          1   4.786379 |
|   30    5   0   0   5.193888   0      2       1          2   4.581161 |
|   30    6   0   0   4.856616   1      3      0          2   5.193888 |
|   30    7   0   0   5.479577   0      3      1          3   4.856616 |
|----------------------------------------------------------------------|
```

Subject 1 died between visits 1 and 2. Subject 2 remained AIDS-free to the end of follow-up. Subject 9 developed AIDS between visits 5 and 6. Subject 30 was lost to follow-up after visit 7.

4.1.2 The command

The g-computation procedure was applied to the first dataset using the following command:

```
gformula y a l a_lag l_lag cuma cuma_lag id t, out(y) com(y:logit, ///
 l:regress, a:logit) eq(y:l_lag cuma_lag, l:l_lag a_lag, a:l a_lag) ///
 id(id) t(t) var(l) intvars(a) interventions(a=1 if t<10, ///
 a=0 if t<=1 \ a=1 if t>1 & t<10, a=0 if t<=3 \ a=1 if t>3 & t<10, ///
 a=0 if t<=5 \ a=1 if t>5 & t<10, a=0 if t<=7 \ a=1 if t>7 & t<10, ///
 a=0 if t<=9) pooled lag(l_lag a_lag cuma_lag) lagrules(l_lag: l 1, ///
 a_lag: a 1, cuma_lag: cuma 1) msm(stcox cuma_lag) derived(cuma) ///
 derrules(cuma:cuma_lag+a) seed(79)
```

Six static regimes are being compared:

1. $(A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9) = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$
2. $(A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9) = (0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1)$
3. $(A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9) = (0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1)$
4. $(A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9) = (0, 0, 0, 0, 0, 0, 1, 1, 1, 1)$
5. $(A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9) = (0, 0, 0, 0, 0, 0, 0, 0, 1, 1)$
6. $(A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9) = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$

In the second analysis, we use the same dataset, but we compare dynamic regimes. Here is the code:

```
gformula y a l a_lag l_lag cuma cuma_lag id t, out(y) com(y:logit, ///
 l:regress, a:logit) eq(y:l_lag cuma_lag, l:l_lag a_lag, a:l a_lag) ///
 id(id) t(t) var(l) intvars(a) interventions(a=0 if t<10 & l>6.9 ///
 \ a=1 if t<10 & l<=6.9, a=0 if t<10 & l>6.55 \ a=1 if t<10 & l<=6.55, ///
 a=0 if t<10 & l>6.2 \ a=1 if t<10 & l<=6.2, a=0 if t<10 & l>5.3 \ ///
 a=1 if t<10 & l<=5.3, a=0 if t<10 & l>4.6 \ a=1 if t<10 & l<=4.6) ///
 dynamic pooled lag(l_lag a_lag cuma_lag) lagrules(l_lag: l 1, ///
 a_lag: a 1, cuma_lag: cuma 1) derived(cuma) derrules(cuma:cuma_lag+a) ///
 seed(801)
```

The dynamic regimes being compared are all of the type 'treat at time t if and only if $L_t < x$', with x taking the values 6.9, 6.55, 6.2, 5.3 and 4.6 in the five different regimes being compared.

Finally, we analyse the second dataset (with losses to follow-up and censoring due to death), and compare the same six static regimes as listed above using the following code:

```
gformula y a l a_lag l_lag d cuma cuma_lag id t, out(y) com(y:logit, ///
 l:regress, a:logit, d:logit) eq(y:l_lag cuma_lag, l:l_lag a_lag, ///
 a:l a_lag, d:l_lag cuma_lag) id(id) t(t) var(l) intvars(a) ///
 interventions(a=1 if t<10, a=0 if t<=1 \ a=1 if t>1 & t<10, ///
 a=0 if t<=3 \ a=1 if t>3 & t<10, a=0 if t<=5 \ a=1 if t>5 & t<10, ///
 a=0 if t<=7 \ a=1 if t>7 & t<10, a=0 if t<=9) pooled ///
 lag(l_lag a_lag cuma_lag) lagrules(l_lag: l 1, a_lag: a 1, ///
 cuma_lag: cuma 1) msm(stcox cuma_lag) derived(cuma) ///
 derrules(cuma:cuma_lag+a) death(d) seed(79)
```

4.1.3 The output

Here is the (abridged) output from the first analysis (the comparison of static regimes with no losses to follow-up and no deaths):

```
G-computation formula estimates for the parameters of the specified marginal
   structural model

         Specified MSM: stcox cuma_lag

 ------------------------------------------------------------------------------
              |  G-computation
              |   estimate of    Bootstrap                        Normal-based
            y |      Coef.       Std. Err.      z    P>|z|     [95% Conf. Interval]
 -------------+----------------------------------------------------------------
```

```
     cuma_lag |   -.2170718      .0394803    -5.5    0.000    -.2944518   -.1396919
 ----------------------------------------------------------------------------------

G-computation formula estimates of the average log incidence rates under each of
   the specified interventions and under no intervention (i.e. as simulated under
   the observational regime). For comparison, the average log incidence rate in the
   observed data is also shown.

         Specified interventions:
              Intervention 1: a=1 if t<10
              Intervention 2: a=0 if t<=1 \ a=1 if t>1 & t<10
              Intervention 3: a=0 if t<=3 \ a=1 if t>3 & t<10
              Intervention 4: a=0 if t<=5 \ a=1 if t>5 & t<10
              Intervention 5: a=0 if t<=7 \ a=1 if t>7 & t<10
              Intervention 6: a=0 if t<=9

 ----------------------------------------------------------------------------------
              |  G-computation
              |   estimate of    Bootstrap                         Normal-based
            y |   av. log IR     Std. Err.      z     P>|z|     [95% Conf. Interval]
 -------------+--------------------------------------------------------------------
       Int. 1 |   -3.922883      .1207616   -32.48   0.000    -4.159572   -3.686195
       Int. 2 |   -3.451616      .0898725   -38.41   0.000    -3.627763   -3.275469
       Int. 3 |   -3.156249      .0882479   -35.77   0.000    -3.329212   -2.983286
       Int. 4 |   -3.143188      .0938219   -33.5    0.000    -3.327076   -2.959301
       Int. 5 |   -3.045274      .0974953   -31.24   0.000    -3.236361   -2.854187
       Int. 6 |   -2.909873       .102833   -28.3    0.000    -3.111422   -2.708324
 -------------+--------------------------------------------------------------------
  Obs. regime |
    simulated |   -3.244703      .0839446   -38.65   0.000    -3.409231   -3.080175
     observed |   -3.374291
 ----------------------------------------------------------------------------------

G-computation formula estimates of the cumulative incidence under each of the
   specified interventions and under no intervention (i.e. as simulated under the
   observational regime). For comparison, the cumulative incidence in the observed
   data is also shown.

 ----------------------------------------------------------------------------------
              |  G-computation
              |   estimate of    Bootstrap                         Normal-based
            y |  cum. incidence  Std. Err.      z     P>|z|     [95% Conf. Interval]
 -------------+--------------------------------------------------------------------
       Int. 1 |        .174      .0211655     8.22   0.000     .1325164    .2154836
       Int. 2 |        .263      .0203068    12.95   0.000     .2231995    .3028005
       Int. 3 |         .34      .0230262    14.77   0.000     .2948695    .3851305
       Int. 4 |        .349      .0269586    12.95   0.000     .2961622    .4018378
       Int. 5 |         .38      .0299038    12.71   0.000     .3213895    .4386105
```

```
      Int. 6 |        .426       .032928    12.94   0.000     .3614623    .4905377
-------------+----------------------------------------------------------------
 Obs. regime |
   simulated |        .318      .0207657    15.31   0.000        .2773       .3587
    observed |         .28
------------------------------------------------------------------------------
```

All three tables point towards a beneficial effect of treatment: the more treatment a subject receives the longer s/he survives AIDS-free. This is seen from the negative log hazard ratio associated with cumulative treatment (corresponding to a HR of 0.805, 95% CI [0.745,0.870]) from the results of the MSM, and from the increasing average log incidence rates and cumulative incidences seen as we move down the other two tables. 43% of the study participants were simulated as having developed AIDS during the hypothetical study in which treatment was withheld (intervention 6), whereas only 17% were simulated to have developed AIDS when treatment was prescribed at all times.

There is a small difference (31% vs. 28% for the cumulative incidences and −3.24 vs. −3.37 for the average log incidence rates) between the simulated and observed data under the observational regime. However, these differences are small relative to the standard error, and thus there is no cause for concern due to this check.

Here is the (abridged) output from the second analysis, comparing dynamic regimes:

```
G-computation formula estimates of the average log incidence rates under each of
   the specified interventions and under no intervention (i.e. as simulated under
   the observational regime). For comparison, the average log incidence rate in the
   observed data is also shown.

         Specified interventions:
              Intervention 1: a=0 if t<10 & l>6.9 \ a=1 if t<10 & l<=6.9
              Intervention 2: a=0 if t<10 & l>6.55 \ a=1 if t<10 & l<=6.55
              Intervention 3: a=0 if t<10 & l>6.2 \ a=1 if t<10 & l<=6.2
              Intervention 4: a=0 if t<10 & l>5.3 \ a=1 if t<10 & l<=5.3
              Intervention 5: a=0 if t<10 & l>4.6 \ a=1 if t<10 & l<=4.6

------------------------------------------------------------------------------
             |  G-computation
             |   estimate of    Bootstrap                     Normal-based
           y |   av. log IR     Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Int. 1 |   -3.799534      .1200517   -31.65   0.000    -4.034831   -3.564237
      Int. 2 |   -3.603487      .1168704   -30.83   0.000    -3.832549   -3.374426
      Int. 3 |   -3.587729      .1152303   -31.14   0.000    -3.813577   -3.361882
      Int. 4 |   -3.412725       .091166   -37.43   0.000    -3.591407   -3.234043
      Int. 5  |   -3.148477      .0900034   -34.98   0.000    -3.324881   -2.972074
-------------+----------------------------------------------------------------
Obs. regime  |
   simulated  |   -3.344836      .0861595   -38.82   0.000    -3.513706   -3.175967
    observed  |   -3.374291
------------------------------------------------------------------------------
```

```
G-computation formula estimates of the cumulative incidence under each of the
   specified interventions and under no intervention (i.e. as simulated under the
   observational regime). For comparison, the cumulative incidence in the observed
   data is also shown.

 ------------------------------------------------------------------------------
              |  G-computation
              |   estimate of    Bootstrap                        Normal-based
            y |  cum. incidence  Std. Err.      z    P>|z|     [95% Conf. Interval]
 -------------+----------------------------------------------------------------
      Int. 1  |        .194      .0214549     9.04   0.000     .1519491    .2360509
      Int. 2  |         .23      .0214971    10.7    0.000     .1878664    .2721336
      Int. 3  |        .237      .0223084    10.62   0.000     .1932762    .2807238
      Int. 4  |        .276      .0220461    12.52   0.000     .2327903    .3192097
      Int. 5  |        .346      .0251734    13.74   0.000      .296661     .395339
 -------------+----------------------------------------------------------------
 Obs. regime  |
   simulated  |        .295      .0212777    13.86   0.000     .2532966    .3367034
    observed  |         .28
 ------------------------------------------------------------------------------
```

We are not able to estimate the parameters of a marginal structural model from this analysis, as explained above. However, the results from the average log incidence rates and the cumulative incidences confirm that treatment is beneficial, with higher AIDS-free survival achieved under the dynamic regime in which x, the threshold below which ART is administered, is highest. There is very good agreement between the simulated and observed data under the observational regime, and the suggestion is that the observational regime is between regime 4 and regime 5 in terms of AIDS-free survival.

Finally, here is the (abridged) output from the third analysis, with loss to follow-up and censoring due to death:

```
G-computation formula estimates for the parameters of the specified marginal
   structural model

         Specified MSM: stcox cuma_lag
```

```
 ----------------------------------------------------------------------------------
              |  G-computation
              |   estimate of    Bootstrap                                  Normal-based
            y |      Coef.       Std. Err.      z    P>|z|     [95% Conf. Interval]
 -------------+--------------------------------------------------------------------
     cuma_lag |   -.1645945      .0408263    -4.03   0.000    -.2446126   -.0845764
 ----------------------------------------------------------------------------------

G-computation formula estimates of the average log incidence rates under each of
   the specified interventions and under no intervention (i.e. as simulated under
   the observational regime). For comparison, the average log incidence rate in the
   observed data is also shown.

         Specified interventions:
              Intervention 1: a=1 if t<10
              Intervention 2: a=0 if t<=1 \ a=1 if t>1 & t<10
              Intervention 3: a=0 if t<=3 \ a=1 if t>3 & t<10
              Intervention 4: a=0 if t<=5 \ a=1 if t>5 & t<10
              Intervention 5: a=0 if t<=7 \ a=1 if t>7 & t<10
              Intervention 6: a=0 if t<=9

 ----------------------------------------------------------------------------------
              |  G-computation
              |   estimate of    Bootstrap                                  Normal-based
            y |   av. log IR     Std. Err.      z    P>|z|     [95% Conf. Interval]
 -------------+--------------------------------------------------------------------
      Int. 1  |   -3.611751      .1323961   -27.28   0.000    -3.871243   -3.352259
      Int. 2  |   -3.317294      .0903419   -36.72   0.000    -3.494361   -3.140228
      Int. 3  |   -3.161453      .0861944   -36.68   0.000    -3.330391   -2.992515
      Int. 4  |   -2.967694      .0935757   -31.71   0.000    -3.151099   -2.784289
      Int. 5  |   -2.976457      .1048993   -28.37   0.000    -3.182056   -2.770858
      Int. 6  |   -2.846718      .1063066   -26.78   0.000    -3.055075   -2.638361
 -------------+--------------------------------------------------------------------
 Obs. regime  |
   simulated  |   -3.218749      .0875509   -36.76   0.000    -3.390346   -3.047153
    observed  |   -3.467099
 ----------------------------------------------------------------------------------

G-computation formula estimates of the cumulative incidence under each of the
   specified interventions and under no intervention (i.e. as simulated under the
   observational regime). For comparison, the cumulative incidence in the observed
   data is also shown.

 ----------------------------------------------------------------------------------
              |  G-computation
              |   estimate of    Bootstrap                                  Normal-based
            y |  cum. incidence  Std. Err.      z    P>|z|     [95% Conf. Interval]
 -------------+--------------------------------------------------------------------
```

```
  Int. 1 (o)  |        .227      .0263101     8.63   0.000     .1754331    .2785669
         (d)  |        .023      .0100621     2.29   0.022     .0032787    .0427213
  Int. 2 (o)  |        .292      .0212259    13.76   0.000      .250398     .333602
         (d)  |        .027      .0100535     2.69   0.007     .0072955    .0467045
  Int. 3 (o)  |        .329      .0220347    14.93   0.000     .2858128    .3721872
         (d)  |        .049      .0116972     4.19   0.000     .0260739    .0719261
  Int. 4 (o)  |        .387      .0255554    15.14   0.000     .3369122    .4370878
         (d)  |        .067      .0145786     4.6    0.000     .0384264    .0955736
  Int. 5 (o)  |        .385      .0303132    12.7    0.000     .3255871    .4444129
         (d)  |        .081      .0164226     4.93   0.000     .0488123    .1131877
  Int. 6 (o)  |        .424         .0318    13.33   0.000     .3616732    .4863268
         (d)  |        .086      .0178607     4.82   0.000     .0509937    .1210063
 -------------+------------------------------------------------------------------
 Obs. regime  |
 simulated (o)|        .316      .0217485    14.53   0.000     .2733737    .3586263
           (d)|        .051       .010469     4.87   0.000     .0304811    .0715189
 observed (o) |        .252
          (d) |        .042
          (l) |        .194
 --------------------------------------------------------------------------------
 Key: (o) = outcome, (d) = death, (l) = lost to follow-up
```

The conclusions from this analysis are similar, but interpretation is now trickier, due to the fact that death is seen as a competing event. It is also more difficult to compare the simulated and observed data under the observational regime, since the former does not include any losses to follow-up, whereas the latter does.

4.1.4 Comparison with standard analysis

We show below the standard Cox regression analysis for AIDS-free survival given the cumulative treatment, with and without adjusting for the time-varying confounder log(CD4). These are the results for the first simulated dataset (without censoring due to death/loss to follow-up).

```
. stcox cuma_lag

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cuma_lag |   1.253736   .0901163     3.15   0.002     1.088988    1.443407
------------------------------------------------------------------------------


. stcox cuma_lag l
```

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cuma_lag |   1.149773   .1142843     1.40   0.160      .946248    1.397073
           l |   1.191835   .1339614     1.56   0.118     .9561846    1.485562
------------------------------------------------------------------------------
```

Both analyses suggest a *harmful* effect of treatment (although the evidence for this effect is very weak in the adjusted analysis). We know from the way that we simulated the data that treatment is beneficial, and this was confirmed by the g-computation analyses. The unadjusted analysis above is biased since it fails to take into account that the treated subjects at any given visit are less healthy than the untreated subjects (since the decision of whether to treat depends on CD4 count at that visit). However, adjusting for log(CD4), also wrongly suggests that treatment is harmful, since conditioning on future CD4 count masks much of the beneficial effect of the treatment.

A similar picture (in fact, more extreme) is seen when performing the standard analyses on the second dataset (with censoring due to loss to follow-up and death).

```
. stcox cuma_lag

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cuma_lag |   1.396623   .1046369     4.46   0.000     1.205885     1.61753
------------------------------------------------------------------------------

. stcox cuma_lag l

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cuma_lag |   1.473824   .1464632     3.90   0.000     1.212987     1.79075
           l |   .9069914   .1055528    -0.84   0.402     .7220096    1.139366
------------------------------------------------------------------------------
```

4.2 Example II: mediation

4.2.1 The data

A dataset comprising 10,000 subjects was simulated according to the description given in section 1.2.5 as follows:

- Socio-economic position (SEP) is generated as 1 (low) for 30% of subjects, 2 (middle) for 50% of subjects, and 3 (high) for the remaining 20%.
- Alcohol intake (A) in units per day is generated as a zero-inflated skewed distribution. A Bernoulli random variable is generated with probability $0.8I\left(SEP=1\right)+0.7I\left(SEP=2\right)+0.9I\left(SEP=3\right)$. If this binary variable is 0, then $A=0$. Otherwise, $\log\left(A\right)$ is taken from a normal distribution with mean $I\left(SEP=1\right)+0.7I\left(SEP=2\right)+1.2I\left(SEP=3\right)$ and variance 0.25.
- Body mass index $\log\left(BMI\right)$ is generated from a normal distribution with mean $23+I\left(SEP=1\right)+0.4A$ and variance 4.
- The logarithm of GGT (measured in grams per litre) is generated from a normal distribution with mean $2.5+0.02BMI+0.1A$ and variance 1.
- Finally, systolic blood pressure (SBP), measured in mmHg, is generated from a normal distribution with mean $80+0.5BMI+6A+7\log\left(GGT\right)-\log\left(GGT\right)A-5\left(SEP-3\right)$ and variance 100.

Independently and completely at random, 5% of subjects have the alcohol variable missing, 5% have the BMI variable missing and 5% have the GGT variable missing. As a result, 8,620 subjects have complete data, 442 have GGT only missing, 424 have BMI only missing, and 427 have alcohol only missing. A further 26 subjects are missing both alcohol and BMI (but have GGT observed), 26 are missing both alcohol and GGT (but have BMI observed), and 34 are missing both BMI and GGT (but have alcohol observed). Finally, 1 subject has a missing value for all three variables.

The data for the first three subjects in the dataset (all with $SEP=1$) are shown below. `log_ggt~c` is an abbreviation of `log_ggt_alc`, the product of `log_ggt` and `alc`, `alc_sbp` is the product of `alc` and `sbp`, and `log_ggt~p` is an abbreviation of `log_ggt_sbp`, the product of `log_ggt` and `sbp`.

```
+-----------------------------------------------------------------------------------+
| sep        alc        bmi    log_ggt        sbp  log_ggt~c    alc_sbp   log_ggt~p |
|-----------------------------------------------------------------------------------|
|   1   1.694112   25.23217   3.055158   128.4518    5.17578   217.6117    392.4406 |
|   1   4.111055          .           .   136.6855          .   561.9217           . |
|   1   .9138322   22.54041    2.998993   126.3068    2.740576   115.4232    378.7932 |
```

4.2.2 The command

The g-computation procedure was applied using the following command:

```
gformula sep alc bmi log_ggt sbp log_ggt_alc alc_sbp log_ggt_sbp, ///
mediation out(sbp) eq(bmi:i.sep alc, log_ggt:bmi alc, sbp:bmi alc ///
log_ggt log_ggt_alc i.sep) com(bmi:regress, log_ggt:regress, sbp:regress) ///
ex(alc) mediator(log_ggt) control(log_ggt:3) baseline(alc:0) ///
post_confs(bmi) base_confs(sep) derived(log_ggt_alc alc_sbp log_ggt_sbp) ///
derrules(log_ggt_alc:log_ggt*alc, alc_sbp:alc*sbp, log_ggt_sbp:log_ggt*sbp) ///
impute(alc bmi log_ggt) imp_cmd(alc:regress, bmi:regress, log_ggt:regress) ///
imp_eq(alc:i.sep bmi log_ggt sbp log_ggt_sbp, bmi:i.sep alc log_ggt sbp ///
log_ggt_alc, log_ggt:i.sep alc bmi sbp alc_sbp) seed(79)
```

4.2.3 The output

Here is the (abridged) output:

```
G-computation formula estimates of the total causal effect, the natural
   direct/indirect effects, and the controlled direct effect

 --------------------------------------------------------------------------------
             |  G-computation   Bootstrap                        Normal-based
             |    estimate      Std. Err.      z    P>|z|     [95% Conf. Interval]
 -------------+------------------------------------------------------------------
       TCE   |    7.516356      .2206128    34.07   0.000      7.083963    7.948749
       NDE   |    6.389072      .2204688    28.98   0.000      5.956961    6.821183
       NIE   |    1.127283      .1681255     6.71   0.000      .7977635    1.456803
       CDE   |    6.301131      .2068248    30.47   0.000      5.895762      6.7065
 --------------------------------------------------------------------------------
```

The conclusion here is that alcohol intake has a causal effect on systolic blood pressure. If everyone were to stop drinking, the average SBP would fall by 7.51 units (95% CI [7.08,7.95]). Only a small part of this reduction (1.13 units) is mediated through GGT. The majority of the effect is direct, *i.e.* it acts through BMI and other pathways.

4.2.4 Comparison with standard analysis

Here are the standard analyses we might have used on these data, as described in the **Introduction**. We use multiple imputation using chained equations (with the same imputation models as above) to deal with the missing data in a comparable way. 5 proper imputations for each missing value are drawn (using `ice` in Stata) and the results are analysed and combined using the `mim` command. Such multiple proper imputations are now required since we use analytical standard errors, rather than bootstrapping.

```
. xi:mim: regress sbp i.sep alc
i.sep             _Isep_1-3           (naturally coded; _Isep_1 omitted)

Multiple-imputation estimates (regress)                  Imputations =       5
Linear regression                                        Minimum obs =   10000
                                                         Minimum dof =   606.6

------------------------------------------------------------------------------
         sbp |      Coef.  Std. Err.      t    P>|t|    [95% Conf. Int.]     FMI
-------------+----------------------------------------------------------------
     _Isep_2 |  -6.20147    .269765  -22.99    0.000   -6.73084  -5.6721   0.015
     _Isep_3 |  -11.1104    .336718  -33.00    0.000   -11.7712 -10.4497   0.022
         alc |   3.27134    .066665   49.07    0.000    3.14041  3.40226   0.081
       _cons |   123.994    .262807  471.81    0.000    123.478   124.51   0.023
------------------------------------------------------------------------------
```

```
. xi:mim: regress sbp i.sep alc log_ggt bmi
i.sep             _Isep_1-3           (naturally coded; _Isep_1 omitted)

Multiple-imputation estimates (regress)                  Imputations =       5
Linear regression                                        Minimum obs =   10000
                                                         Minimum dof =   177.7

------------------------------------------------------------------------------
         sbp |      Coef.  Std. Err.      t    P>|t|    [95% Conf. Int.]     FMI
-------------+----------------------------------------------------------------
     _Isep_2 |  -5.46916    .251701  -21.73    0.000   -5.96308 -4.97523   0.043
     _Isep_3 |  -10.3605    .311652  -33.24    0.000    -10.972 -9.74891   0.051
         alc |   2.53784    .067243   37.74    0.000    2.40514  2.67053   0.158
     log_ggt |   4.82774    .103892   46.47    0.000    4.62386  5.03161   0.044
         bmi |   .435676    .051427    8.47    0.000    .334759  .536593   0.023
       _cons |   99.0651    1.28187   77.28    0.000    96.5497  101.581   0.015
------------------------------------------------------------------------------
```

These estimates are not directly comparable with those obtained using the g-computation procedure, since the coefficient of alcohol in the analyses above are for a unit change in units consumed per day. Since the average number of units consumed per day in the simulated dataset is 2.22, the equivalent total effect estimated by standard regression would be approximately $3.27 \times 2.2 = 7.25$, that is similar to the 7.51 obtained above, as we would expect. However, the coefficient of alcohol in the second regression analysis, if interpreted naïvely, would be taken to represent the direct effect, not mediated by GGT, with a derived indirect effect of approximately $7.25 - (2.54 \times 2.22) = 1.62$ appearing to be larger from this analysis than from the g-computation analysis. In other words, the standard analysis would lead us to conclude that more of the effect is mediated by GGT than is truly the case. This is to be expected, since some of the direct effect of alcohol on SBP (*i.e.* that which is not mediated by GGT) acts through BMI, and this part of the effect is not correctly apportioned in the standard analysis above, leading to the underestimation of the direct effect.

4.3 A warning on computation time

The `gformula` command is computationally very intensive, and computation time increases exponentially with increasing number of time-points. In the time-varying confounding example above, with $T = 9$, fitting the parametric models, simulating the data under each intervention and then analysing each simulated datset, takes around 30 seconds on a standard PC. Thus, if 1,000 bootstrap samples are required, the whole analysis takes over 8 hours. However, bootstrapping is ideally suited to task-sharing, and the command runs in a fraction of the time on a high-performance computer cluster.

5 Final remarks

In problems concerning time-varying confounding and mediation, we have reiterated that standard regression analyses are invalid when confounders are affected by the exposure. The g-computation procedure is valid under a weaker set of assumptions that allow for confounders to be affected by past exposure. The structural assumption needed for this procedure to be valid is that a sufficient set of confounders have been measured. In addition, the procedure requires that correct parametric models be postulated for the post-baseline variables in the observational data.

Alternative semiparametric models and estimation methods have been proposed (Robins *et al.*, 1992, 2000). These are g-estimation of structural nested models and inverse probability weighted estimation of marginal structural models. These alternative methods rely on fewer parametric modelling assumptions and are therefore less prone to model misspecification bias. In addition, these semiparametric approaches do not require Monte Carlo simulation and are thus computationally less intensive. Their implementation in Stata has been demonstrated (Sterne and Tilling, 2002; Fewell *et al.*, 2004).

However, the g-computation procedure has some clear advantages over these methods: it can more easily deal with complex multivariate (or joint) interventions, and can compare the hypothetical interventions with the observational regime, which can be important in informing policy (Taubman *et al.*, 2009). These advantages are in addition to that of increased statistical efficiency which is gained at the price of stronger modelling assumptions (Daniel *et al.*, 2010).

We believe that the g-computation procedure is a valuable tool in many settings. Although first proposed by Robins in 1986, it has not been very widely used, partly as a result of its apparent complexity and the lack of software routines, until the recent GFORMULA macro in SAS (Taubman *et al.*, 2009). We hope that this Stata routine will help to make this method more accessible to a wider audience of applied researchers.

Acknowledgements

`gformula.ado` makes use of the `detangle` and `formatlist` procedures from `ice.ado`, by kind permission of Patrick Royston. The command is inspired by the GFORMULA macro in SAS (Taubman *et al.*, 2009).

References

Angrist, J. D. and Pischke, J.-S. (2009) *Mostly harmless econometrics: an empiricist's companion.* Princeton University Press.

van Buuren, S., Boshuizen, H. C. and Knook, D. L. (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, **18**, 681–694.

Cain, L. E., Robins, J. M., Lanoy, E., Logan, R., Costagliola, D. and Hernán, M. A. (2010) When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *International Journal of Biostatistics*, **6**, Article 18.

D'Agostino, R. B., Lee, M.-L., Belanger, A. J., Cupples, L. A., Anderson, K. and Kannel, W. B. (1990) Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham heart study. *Statistics in Medicine*, **9**, 1501–1515.

Daniel, R. M., Cousens, S. N., De Stavola, B. L., Kenward, M. G. and Sterne, J. A. C. (2010) Methods for dealing with time-varying confounding. *Statistics in Medicine*, p. *(under review)*.

Didelez, V. (2006) Direct and Indirect Effects of Sequential Treatments. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*. pp. 138–146.

Fewell, Z., Hernán, M., Wolfe, F., Tilling, K., Choi, H. and Sterne, J. (2004) Controlling for time-dependent confounding using marginal structural models. *The Stata Journal*, **4**, 402–420.

Greenland, S., Pearl, J. and Robins, J. M. (1999) Causal diagrams for epidemiological research. *Epidemiology*, **10**, 37–48.

Hafeman, D. (2009) “Proportion explained”: a causal interpretation for standard measures of indirect effect? *American Journal of Epidemiology*, **170**, 1443–1448.

Hernán, M. A., Hernández-Díaz, S. and Robins, J. M. (2004) A structural approach to selection bias. *Epidemiology*, **15**, 615–625.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data.* New York: Wiley.

Morgan, S. L. and Winship, C. (2007) *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge University Press.

Pearl, J. (2001) Direct and indirect effects. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*. *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pp. 411–420.

Pearl, J. (2009) *Causality.* Cambridge University Press, second edition.

Robins, J. M. (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**, 1393–1512.

Robins, J. M. (2003) *Semantics of causal DAG models and the identification of direct and indirect effects. In* Highly Structured Stochastic Systems. NY: Oxford University Press.

Robins, J. M. and Gill, R. D. (1997) Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine*, **16**(1), 39–56.

Robins, J. M. and Greenland, S. (1992) Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143–155.

Robins, J. M. and Hernán, M. A. (2009) *Longitudinal Data Analysis*, chapter 23: Estimation of the causal effects of time-varying exposures, pp. 553–599. New York: Chapman and Hall/CRC Press.

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

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

Rothman, K. J., Greenland, S. and Lash, T. L. (2008) *Modern Epidemiology.* Lippincott Williams & Wilkins, third edition.

Sterne, J. A. C. and Tilling, K. (2002) Gestimation of causal effects, allowing for timevarying confounding. *The Stata Journal*, **2**, 164–182.

Taubman, S. L., Robins, J. M., Mittleman, M. A. and Hernán, 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.

Tsiatis, A. A. (2006) *Semiparametric Theory and Missing Data*. Springer, New York.