The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index®
- Current Contents/Social and Behavioral Sciences®
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch®)
- Scopus™
- Social Sciences Citation Index®

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata, Mata, NetCourse, and Stata Press are registered trademarks of StataCorp LP.

The Stata Journal (2011)
11, Number 4, pp. 479–517

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

Rhian M. Daniel
Centre for Statistical Methodology
London School of Hygiene and Tropical Medicine
London, UK
rhian.daniel@lshtm.ac.uk

Bianca L. De Stavola
Centre for Statistical Methodology
London School of Hygiene and Tropical Medicine
London, UK

Simon N. Cousens
Centre for Statistical Methodology
London School of Hygiene and Tropical Medicine
London, UK

Abstract. This article describes a new command, `gformula`, that is an implementation of the g-computation procedure. It is used to estimate the causal effect of time-varying exposures on an outcome in the presence of time-varying confounders that are themselves also affected by the exposures. The procedure also addresses the related problem of estimating direct and indirect effects when the causal effect of the exposures on an outcome is mediated by intermediate variables, and in particular when confounders of the mediator–outcome relationships are themselves affected by the exposures. A brief overview of the theory and a description of the command and its options are given, and illustrations using two simulated examples are provided.

Keywords: st0238, gformula, causal inference, g-computation formula, time-varying confounding, mediation, direct and indirect effects

1 Introduction

1.1 Time-varying confounding

The setting

Longitudinal studies, where data are collected sequentially at several points in time, are common in many areas of research, including epidemiology, clinical trials, ecology, sociology, and econometrics. 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—that is, at each visit—when the assessment of whether the event has occurred since the last visit is made.

In attempting to estimate this causal effect, it is important to consider the role of confounding variables. Informally, these variables influence both the explanatory variable and the outcome (more details are presented below). Failure to consider the role of confounding variables 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, Greenland, and Lash 2008; Morgan and Winship 2007; Angrist and Pischke 2009). This article focuses on the specific problem of time-varying confounders, that is, factors that potentially confound the causal relationship between a 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 (that is, regression adjustment) for dealing with confounding can no longer be applied (Robins 1986; Robins and Hernán 2009; Daniel et al. 2011).

In this article, we describe the g-computation procedure and its implementation in a newly written Stata command. This procedure was first suggested by Robins (1986) to overcome the limitations of standard methods. Alternative estimators have also been suggested by Robins and his colleagues, and these are discussed further in section 5.

The scenario to be considered is best illustrated using a causal diagram (Greenland, Pearl, and Robins 1999) such as the one depicted in figure 1. The arrows in this diagram represent the assumed direction of causal influence. $\{A_0, A_1, \ldots, A_T\}$ represent the explanatory variables of interest measured at time points $0, 1, \ldots, T$. $\{L_0, L_1, \ldots, L_T\}$ represent the potential confounders 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 (at visit $T+1$), and U is a set of unmeasured factors that influence $\{L_0, L_1, \ldots, L_T\}$ and Y.

Notice that there are no arrows from U to $\{A_0, A_1, \ldots, A_T\}$ and no other common causes (say, V) of $\{A_0, A_1, \ldots, A_T\}$ 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 would be independent of Y. This is known as the "no unmeasured confounders" assumption; it means that at each visit t, a sufficient set of confounders of the relationship between A_t and Y is measured. Also notice the arrows from A_t to L_{t+1}; these represent that the confounder at one time point may be influenced by the explanatory variable at the previous time point. We could have included arrows from A_0 to L_2, from L_0 to A_1, from L_0 to Y, etc. These were left out simply to make the diagram more readable, but the omission of the arrows from U to $\{A_0, A_1, \ldots, A_T\}$ is crucial.

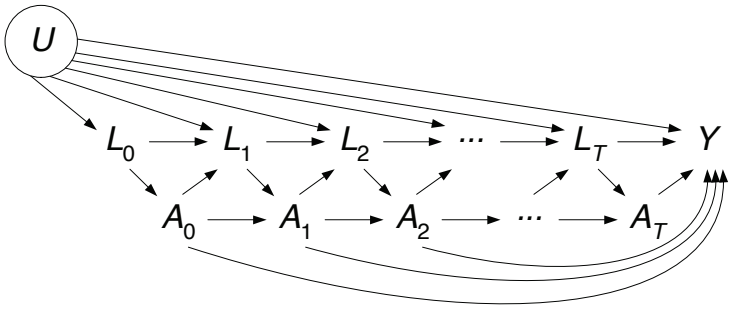

Figure 1. A causal diagram depicting time-varying confounders affected by exposure, when the outcome is measured at the end of follow-up

If instead the outcome is time to event, measured at the discrete time points $1, 2, \ldots, T+1$, then the appropriate causal diagram is the one shown in figure 2, where each Y_t is a binary variable signifying whether 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 $\{A_0, A_1, \ldots, A_T\}$.

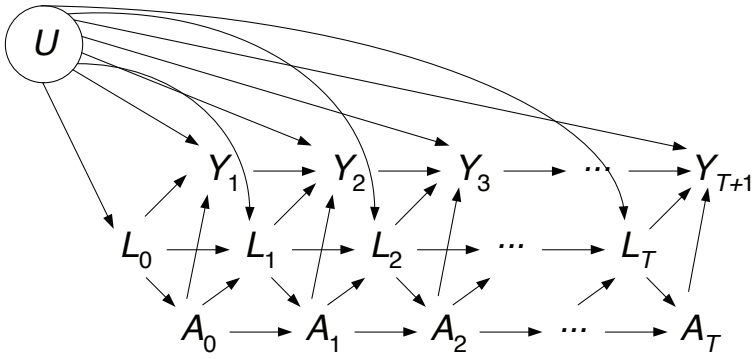

Figure 2. A causal diagram depicting time-varying confounders affected by exposure when the outcome is time to event

Limitation of standard methods

The standard method to adjust for confounding due to L is, in a regression analysis, to condition on $\{L_0, L_1, \ldots, L_T\}$. 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, Pearl, and Robins 1999) from A to Y, allowing us to estimate consistently the joint causal effect

(conditional on $\{L_0, L_1, \ldots, L_T\}$) of $\{A_0, A_1, \ldots, A_T\}$ on Y (or $\{Y_1, Y_2, \ldots, Y_{T+1}\}$ in the case of a time-to-event outcome).

However, in the situation depicted by figures 1 and 2, where the confounder is influenced by past values of the explanatory variables, conditioning on $\{L_0, L_1, \ldots, L_T\}$ is not valid for two reasons. Consider, for example, the causal effect of A_0 on Y (figure 1). By conditioning on L_0, we have successfully blocked the backdoor (noncausal) 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 $A_0 \rightarrow L_1 \rightarrow L_2 \rightarrow \cdots \rightarrow Y$ (and many others), which represents part of the causal effect of A_0 on Y. Furthermore, because U and A_0 both influence L_1, conditioning on L_1 induces a noncausal association between A_0 and U (see Greenland, Pearl, and Robins [1999]), thereby opening up a new backdoor path from A_0 through U to Y. This is called "collider-stratification bias" (Hernán, Hernández-Díaz, and Robins 2004) and applies similarly to figure 2.

Example I

In a longitudinal study of antiretroviral therapy (ART) in HIV research, A_t is a binary variable indicating whether a subject is prescribed ART at time point t, L_t is the CD4 count measured at time t, and Y_t is a binary variable indicating whether the subject develops AIDS in the interval $(t-1, t]$. In an observational study, we would expect the decision as to whether 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 count 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

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.

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. The extent to which the coefficient of X changes between models (i) and (ii) is often 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 there are confounders L of the M–Y relationship (as shown in figure 3), because 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. However, if L is affected by X (as shown in figure 3), then conditioning on L also blocks the path $X \rightarrow L \rightarrow Y$, which is part of the direct effect of X on Y (because 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 interaction between X and M in their effect on Y.

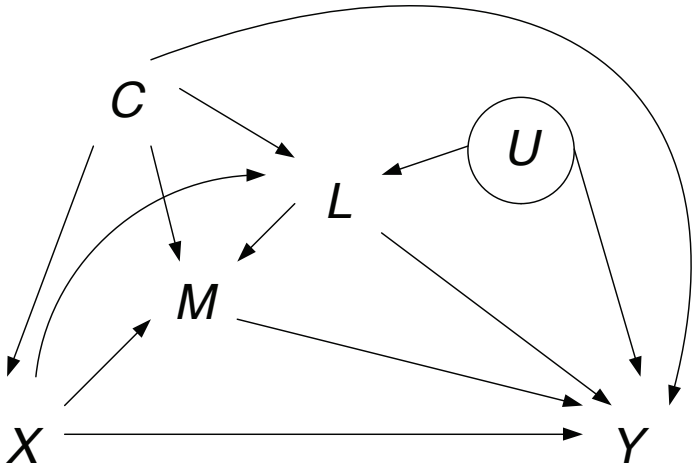

Figure 3. A causal diagram depicting mediation with mediator–outcome confounders affected by the exposure

Relationship to time-varying confounding

To see the link between the two settings, note that figure 3 is the same as figure 1 with $T = 1$, $L_0 = C$, $A_0 = X$, $A_1 = M$, and $L_1 = L$. Thus in the time-varying confounding example, when we talk about the joint causal effect of $\{A_0, A_1, \ldots, A_T\}$ on Y, we are more specifically talking about a collection of $T + 1$ direct causal effects: the direct effect of A_0 on Y not mediated by $\{A_1, A_2, \ldots, A_T\}$, the direct effect of A_1 on Y not mediated by $\{A_2, A_3, \ldots, A_T\}$, and so on.

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 while keeping m fixed. For example, if X is univariate and binary, we might specifically consider the CDE_m to be

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

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

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

It would be desirable to use these quantities to infer an indirect effect as the difference between the total effect and the direct effect, but this is problematic because the CDE is a function of m. If there is no effect of X on M, then there would be no mediation and we would expect the indirect effect to be zero, and hence the total and direct effects to be equal. But this cannot be the case with the CDE because it generally depends on m, whereas the total effect does not.

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

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

We see immediately that in the absence of an effect of X on M, $M(0) = M(1)$ and the NDE is equal to the TCE, as desired.

Then the natural indirect effect (NIE_{x_1}) can be defined as the difference between the TCE and the NDE. Thus it compares $E[Y\{x_1, M(x)\}]$ for different values of x, while keeping x_1 fixed (at a natural choice of "nonbaseline" value). This is best illustrated by thinking again of a binary X, when the NIE becomes

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

There has been much discussion and controversy in recent literature over the definitions, interpretations, and assumptions required to estimate NDEs and NIEs (Robins 2003; Didelez, Dawid, and Geneletti 2006; Petersen, Sinisi, and van der Laan 2006; Hafeman 2009; Robins and Richardson 2010). However, these concerns are beyond the scope of this article.

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, gamma-glutamyl transpeptidase (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. Body mass index (BMI) is thought to affect both GGT and SBP, and socioeconomic 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 figure 3, with $X =$ alcohol intake, $M =$ GGT, $Y =$ SBP, $L =$ 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, 1987a,b, 1989b) and further discussed by Robins (1989a, 1997), Robins, Greenland, and Hu (1999), Gill and Robins (2001), Robins and Hernán (2009), Taubman et al. (2009), and Daniel et al. (2011).

In the next section, we give a brief overview of the g-computation procedure. Then we describe in section 5 our implementation of it using a new command (`gformula`) in Stata. In section 4, we illustrate the command 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

The basic idea

The g-computation procedure works by first modeling 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 $\{A_0, A_1, \ldots, A_T\}$ been determined by intervention rather than been allowed to evolve naturally as in the observational data. The modeling and simulation is carried out forward in time; that is, we start by modeling 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 to simulate the data at time 2 under the various interventions (on time 0 and time 1 exposures), and so on. All postbaseline confounders and outcomes are simulated under each intervention. We can then pursue causal inference by comparing the outcomes under different interventions as if these had been generated from a randomized experiment.

Fitting the models

Univariate L_t ($t \geq 1$)
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 $f_{L_1|L_0,A_0}\left(l_1\,|l_0,a_0;\alpha_1\right)$ is the probability density function from this model. If L_1 is discrete, then $f_{L_1|L_0,A_0}\left(l_1\,|l_0,a_0;\alpha_1\right)$ is a conditional probability.

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

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

When the outcome is time to event—that is, 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, ..., L_{t-1}, A_{t-1} and $Y_{t-1} = 0$ is specified. The estimates $\widehat{\beta}_t$ of the parameters β_t in $f_{Y_t|L_0,A_0,\ldots,L_{t-1},A_{t-1}}\left(y_t\,|l_0,a_0,\ldots,l_{t-1},a_{t-1};\beta_t\right)$ are obtained from the subset of the observational data with $Y_{t-1} = 0$ (that is, those still in the risk set).

Presently, only `regress`, `logit`, `mlogit`, and `ologit` are supported by `gformula` in the fitting of these models. There is an option either to fit the models separately at each time point (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 5.

Multivariate L_t ($t \geq 1$)
The description above assumes that each L_t is univariate. If L_t contains $q_t > 1$ variables, then these must be ordered $L_t = (L_{t,1},L_{t,2},\ldots,L_{t,q_t})$, for example, and the model for $L_{t,j}$ should be specified conditional on $L_{t,1}$, ..., $L_{t,j-1}$ as well as on $L_0,\ldots,L_{t-1}$ and $A_0,\ldots,A_{t-1}$. The order can be chosen to reflect the hypothesized causal ordering of the q_t components of L_t, but this is not strictly necessary; the important thing is that the simulated values of $(L_{t,1},L_{t,2},\ldots,L_{t,q_t})$ should be correlated in a way that matches the correlation seen in the observational data. For simplicity, we assume henceforth that each L_t is univariate, but this assumption is not required.

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, that is, under the intervention $A_0 = 0$, $A_1 = 0$, ..., $A_T = 0$.

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 $\{A_0, A_1, \ldots, A_T\}$. Thus $L_0^* = L_0$.

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

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

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

Thus we have simulated all postbaseline variables, including the potential outcome (given the no unmeasured confounders assumption and the modeling assumptions made during the model-fitting stage) under the hypothetical intervention in which treatment is withheld from all subjects at all times. At each stage, the conditional density used for simulation only conditions on past values of the exposure and confounder. Because U is unmeasured, the simulation is done marginally over the unobserved distribution of U, but because U is not a confounder of the A–Y relationships, this does not introduce bias (Daniel et al. 2011).

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

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 $\{A_0, A_1, \ldots, A_T\}$ with 1 rather than 0. Similarly, many more hypothetical interventions can be compared using this procedure. For example, we could simulate under the intervention "treat at alternate time points" or "treat after time point 3". When $\{A_0, A_1, \ldots, A_T\}$ are continuous, different hypothetical interventions that set A_0 to a_0, …, A_T to a_T can

be compared for different combinations of values of $\{a_0, a_1, \ldots, a_T\}$. This can be extended to multivariate exposures when each of $\{A_0, A_1, \ldots, A_T\}$ and $\{a_0, a_1, \ldots, a_T\}$ are vectors.

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. Because 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 standardization 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 can be plotted for the different interventions. Again, because 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 modeling 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, and Kaplan–Meier curves can be attributed to the causal effect of the exposure.

The number of subjects simulated

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 this increases the Monte Carlo simulation error and hence decreases precision. This is not recommended unless the original dataset is very large, causing the computation time to be unacceptably large.

Standard errors and inference using the bootstrap

Standard errors and confidence intervals (CIs) 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 bootstrap sample.

Comparing dynamic regimes

The interventions considered so far are all termed static, because (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 behavior 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 prespecified 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 (for example, in the HIV study, the optimal regime might be defined as the regime that maximizes 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^*, ..., A_{t-1}^*, L_t^*.

Suppose that for a particular subject, $L_0^* > x$,..., $L_{t-1}^* > x$, 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, because given past values of the confounder and exposure, the rule defining the dynamic regime assigns a value to the current exposure with probability 1. See below (the section *Simulating under the observational regime*) for an example of a stochastic dynamic regime. Dynamic regimes can be compared with `gformula` using the `dynamic` option (see section 3.3).

For more details on the comparison of dynamic regimes using the g-computation procedure, see Murphy (2003), Robins and Hernán (2009), and Daniel, De Stavola, and Cousens (forthcoming).

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. Indeed, this is essentially what it is. However, we may need a more parsimonious way of summarizing the comparison. This can be done using a marginal structural model (MSM) (Robins, Hernán, and Brumback 2000).

An 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 $Y(a_0, a_1, \ldots, a_T)$ 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

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

Thus, given the assumptions made by this MSM, the causal effect of $\{A_0, A_1 \ldots, A_T\}$ on Y can be summarized in terms of one parameter (ϕ), representing the cumulative effect of treatment, 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.

In the case of a time-to-event outcome, if the parameters of the model for $\{Y_1, Y_2, \ldots, Y_{T+1}\}$ 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 an MSM using `stcox`.

Presently, only MSMs fit using `regress`, `logit`, and `stcox` are supported by the `gformula` command (using the `msm()` option).

Standard errors and CIs are again obtained by bootstrapping.

According to this definition of an MSM, only static regimes can be compared, and thus the `msm()` option cannot be specified together with `dynamic` in `gformula`. Dynamic MSMs have been developed (Cain et al. 2010) but are currently not supported by `gformula`.

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)—that is, 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 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.

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 above for AIDS-free survival. First, the question is asked, Did this subject survive the interval $(t-1, t]$? Then, conditional on the answer to this being simulated as "yes", the second question is asked: Did the subject develop AIDS in the interval $(t-1, t]$? And the answer is simulated.

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 nonmonotone) missingness; that is, some subjects miss a visit but return at a subsequent visit. In addition, a subject may have a missing value for a subset of the variables at time point t while 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 nonmonotone patterns of missingness [Robins and Gill 1997]), such nonmonotone patterns of missingness can be dealt with via the method of multiple imputation using chained equations (van Buuren, Boshuizen, and Knook 1999). The method as described by van Buuren, Boshuizen, and Knook draws multiple proper imputations in the sense described by Little and Rubin (2002). The imputations are drawn from the distribution of the missing data given the observed. The distinction between proper and improper imputations concerns the parameters of this distribution; in improper imputation, the parameters are replaced by their maximum likelihood estimates, whereas proper imputation uses Bayesian draws from the posterior distributions of these parameters, with different draws used 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.

Because 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, Boshuizen, and Knook (1999) except that we draw only one imputation for each missing value and that imputation is improper. This has been shown to be a valid approach (and indeed superior in terms of efficiency) when Rubin's variance estimator is not being used (Tsiatis 2006, chap. 14).

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

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 that led to 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 $\{A_1^*, A_2^*, \ldots, A_T^*\}$ under this regime are simulated analogously to what was described for $\{L_1^*, L_2^*, \ldots, L_T^*\}$ above. More precisely, for each $t \in [1, T]$, a model for the conditional distribution of L_t given L_0, A_0, $\ldots$, L_{t-1}, A_{t-1} is specified as before, and the estimates $\widehat{\alpha}_t$ of the parameters α_t from the density/probability $f_{L_t|L_0,A_0,\ldots,L_{t-1},A_{t-1}}(l_t|l_0, a_0, \ldots, l_{t-1}, a_{t-1}; \alpha_t)$ are obtained from the observational data as before. But now, in addition, a model for the conditional distribution of A_t given L_0, A_0, $\ldots$, L_{t-1}, A_{t-1}, L_t is specified, and the estimates $\widehat{\alpha}'_t$ of the parameters α'_t from the density/probability $f_{A_t|L_0,A_0,\ldots,L_{t-1},A_{t-1},L_t}$ $(a_t|l_0, a_0, \ldots, l_{t-1}, a_{t-1}, l_t; \alpha'_t)$ are also obtained from the observational data. The data are then simulated in the order $L_1^*, A_1^*, \ldots, L_T^*, A_T^*, Y^*$ as described previously, but with $\{A_1^*, A_2^*, \ldots, A_T^*\}$ replacing $\{A_1, A_2, \ldots, A_T\}$ instead of $\{a_1, a_2, \ldots, a_T\}$, and with each A_t^* being drawn from $f_{A_t|L_0,A_0,\ldots,L_{t-1},A_{t-1},L_T}$ $(a_t|L_0^*, A_0^*, \ldots, L_{t-1}^*, A_{t-1}^*, L_t^*; \widehat{\alpha}'_t)$.

In the absence of 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 modeling assumptions, or the no unmeasured confounding assumption, do not hold. But 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, because 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 (Hubbard and van der Laan 2008; Taubman et al. 2009).

2.2 Mediation

The g-computation procedure for the mediation example works in a similar way except that for the NDEs and NIEs, simulations under different hypothetical interventions need to be combined and additional assumptions are needed for this. 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 NDE), X is set to 1 at the same time as M is set to the simulated value under the intervention $X = 0$, that is, $M^*(0)$.

If X is not binary or if X is multivariate, there may not be a natural comparison (such as 1 versus 0) for calculating the TCE, CDE, or NDEs or NIEs. In this case, the formulas in section 1.2 are replaced with

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

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

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

and

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

where 0 is still the "baseline" value 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 (Hubbard and van der Laan 2008).

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

For the NDE and NIE to be consistently estimated, in addition to the no unmeasured confounding assumption (for the mediator–outcome and exposure–outcome relationships) and the parametric modeling assumptions made in the g-computation procedure,

we also need a no unmeasured confounding assumption for the exposure–mediator relationship and one further assumption. Much has been written about the nature of this additional assumption, and different authors have chosen different assumptions (Hafeman and VanderWeele 2011).

Briefly, the additional assumption either requires that there be no intermediate confounding (that is, that the mediator–outcome confounders are not affected by the exposure) (Pearl 2001) or requires some form of "no interaction" (Robins and Greenland 1992; Petersen, Sinisi, and van der Laan 2006). The `gformula` command will give consistent estimation under any of these assumptions, but because the g-computation formula is specifically used to deal with the problem of intermediate confounding, the most useful additional assumption takes the form of a no interaction assumption. To date, the weakest but sufficient such no interaction assumption is the one proposed by Petersen, Sinisi, and van der Laan (2006), namely, that

$$E\{Y(x,m) - Y(0,m) \,|M(0) = m, C, L\} = E\{Y(x,m) - Y(0,m) \,|C, L\}$$

This assumption states that conditional on baseline and postbaseline confounders, knowing what an individual's mediator would have been under the baseline level of the exposure does not provide any additional information about the CDE.

Such an additional assumption is not required when estimating the CDE alone.

3 The gformula command

3.1 Syntax

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

where *mainvarlist* contains all the variables to be used by the command. 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 in *mainvarlist* using only their names (for example, `agecat`) and without using the prefix "`i.`" (for example, `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**); that is, 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 figure 2. No records should be included in the dataset for subjects who have been censored before that time due to death, loss to follow-up, or (in the case of a time-to-event outcome) having experienced the event before that time.

Any value that is missing and to be imputed using the `impute()` option and its suboptions (including those values at intermittent missing visits, for which a record must be included) should be denoted by a period (`.`) according to Stata's convention for generic missing values. All records containing a missing value for a relevant variable (a variable involved in the analysis) that is not included in the `impute()` option will be dropped, and a completers-only analysis will be performed with respect to such variables.

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

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

3.3 Options

Time-varying confounding options

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

`commands(`*string*`)` specifies which command (`regress`, `logit`, `mlogit`, or `ologit`) 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()` is required.

Commands should be specified for the models for the outcome variable, time-varying confounders, and the time-varying exposure (needed for simulation under the observational regime). 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 (including the variable representing death, if there is censoring due to death), `logit` should be the chosen command because the time to the event (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). `equations()` is required.

Because 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. That is, one equation for L is given, and it is assumed to hold for each $\{L_t : t = 1, \ldots, T\}$.

Variables that are to be treated as categorical variables on the right-hand side 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 or visit.

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

Derived variables should not be included because these will be simulated as a function of the variables from which they are derived.

The variables should be listed in the order described in the section *Fitting the models: Multivariate L_t*, above.

`intvars(`*varlist*`)` specifies that the variables in *varlist* are the variables on which interventions are to be specified. If lagged versions of these variables are to be used, then 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 backward slash (\) (see the example syntax in section 4).

`dynamic` specifies that at least one of the regimes to be compared (other than the observational regime) is dynamic. If this option is not specified, then 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, then it is assumed that the outcome is time to event.

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

`monotreat` specifies that the time-varying exposure in the observational data is binary and that it changes at most once (from zero to one). Thus the exposure data for a given subject consists of a sequence of zeros followed by a sequence of ones (or a

sequence containing only zeros or only ones). This is common in many settings in which treatment may be initiated at some point but never discontinued. Specifying this option affects the way in which data are simulated only under the observational regime. When using the `monotreat` option, the corresponding command for simulating the time-varying exposure should be specified as `logit`.

`death(`*varname*`)` gives the name of the variable (a sequence of binary variables at each time point) that 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.

All censoring (before the final visit) where the variable denoting the death process takes the value 0 is assumed to be 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, then 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 that 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.

Only derived variables based on the time-varying confounders need to be specified, because these need to be simulated. Interactions, say, between time-fixed confounders can simply be included themselves in the list of time-fixed confounders.

`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`, then the code is `derrules(al:a*l)` (and `al` should be included in `derived()` 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 MSM, 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 (`regress`, `logit`, `mlogit`, or `ologit`) 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 right-hand side 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 `imp_cycles(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 set to be smaller.

`samples(#)` specifies the number of bootstrap samples. The default is `samples(1000)`.

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

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

`graph` specifies that a Kaplan–Meier plot of the survival curves under each intervention be displayed. This option is relevant only 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`, that 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 `saving()` option already exists, then it should be overwritten.

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. For a mediation analysis, `mediation` is required.

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

`commands(`*string*`)` specifies which command (`regress`, `logit`, `mlogit`, or `ologit`) 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). `commands()` is required.

Models must be specified for the mediators, the outcome, and the postbaseline 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). `equations()` is required.

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

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

Only derived variables based on the postbaseline variables need to be specified, because these need to be simulated. Interactions, say, between baseline confounders can simply be included themselves in the list of baseline confounders.

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

`msm(`*string*`)` specifies the form of an MSM to be fit, for example, `msm(regress y x m)` or `msm(logit y x m xm)`. Only `regress`, `logit`, and `stcox` are supported at present.

`exposure(`*varlist*`)` specifies the exposure variables.

`mediator(`*varlist*`)` specifies the mediator variables.

`control(`*string*`)` specifies the values at which the mediators should be controlled when estimating the CDE (see the example syntax in section 4). If `control()` is not specified, then only NDEs and NIEs are estimated.

`baseline(`*string*`)` specifies the values of the exposures to be taken as baseline values (see the example syntax in section 4).

`obe` specifies that there is only one binary exposure and that the comparisons should be made between $X = 1$ and $X = 0$. If neither this nor `oce` (see next option) is specified, then comparisons are made between the distribution of X in the observed data and the baseline values.

`oce` specifies that there is only one categorical exposure and that the comparisons should be made between each nonbaseline level of X and the baseline level, as specified using the `baseline()` option above. If neither this nor `obe` is specified, comparisons are made between the distribution of X in the observed data and the baseline values.

`base_confs(`*varlist*`)` specifies the confounders of the exposure–outcome relationships, and—for estimating the NDE and NIE—the confounders of the exposure–mediator re-

lationships. Any mediator–outcome confounders not affected by the exposure should also be listed here.

`post_confs(`*varlist*`)` specifies the confounders of the mediator–outcome relationships that are affected by the exposure. These should be specified in the order described in the section *Fitting the models*: *Multivariate* L_t.

Derived variables should not be included because these will be simulated as a function of the variables from which they are derived.

`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 (`regress`, `logit`, `mlogit`, or `ologit`) 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 right-hand side 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 `imp_cycles(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 `samples(1000)`.

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

`all` specifies that all bootstrap CIs are to be displayed (normal, percentile, bias corrected, and bias corrected and accelerated). The default is to give only normal-based bootstrap CIs. 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 `saving()` option already exists, then it should be overwritten.

4 Illustration using two simulated examples

4.1 Example I: Time-varying confounding

The data

Two datasets are simulated with $T = 9$, according to the description given in section 1.1. In the first dataset, $\{A_0, A_1, \ldots, A_9\}$ are binary treatment variables with $A_t = 1$ if a subject is prescribed ART at visit t and $A_t = 0$ otherwise. $\{L_0, L_1, \ldots, L_9\}$ are the values of the logarithm of CD4 count at each visit. $\{Y_1, Y_2, \ldots, Y_{10}\}$ are binary variables, where $Y_t = 1$ if a subject develops AIDS during the time-interval $(t-1, t]$ and $Y_t = 0$ 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 $t + 1$ onward.

Here are the data for the first three subjects in the first dataset:

```
. use tvc1
. list id t y l a cuma a_lag cuma_lag l_lag if id<4, sepby(id)
```

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	5.950063	0	2	1	2	5.524413
1	3	0	5.230726	1	3	0	2	5.950063
1	4	0	5.624882	0	3	1	3	5.230726
1	5	0	4.959467	1	4	0	3	5.624882
1	6	1	5.496145	1	5	1	4	4.959467
2	0	.	4.686166	0	0	0	0	0
2	1	0	4.05956	0	0	0	0	4.686166
2	2	1	3.423921	1	1	0	0	4.05956
3	0	.	6.051494	0	0	0	0	0
3	1	0	5.407419	0	0	0	0	6.051494
3	2	0	4.752249	1	1	0	0	5.407419
3	3	0	5.155554	1	2	1	1	4.752249
3	4	0	5.670463	0	2	1	2	5.155554
3	5	0	5.168072	1	3	0	2	5.670463
3	6	0	5.551989	1	4	1	3	5.168072
3	7	0	6.211178	0	4	1	4	5.551989
3	8	0	5.481304	0	4	0	4	6.211178
3	9	0	4.899148	0	4	0	4	5.481304
3	10	0	.	.	.	0	4	4.899148

Subject 3 remained AIDS-free until the end of follow-up. Subject 1 developed AIDS between times 5 and 6, and subject 2 between times 1 and 2.

The variable `a_lag` is the lagged version of `a`—that is, it contains the previous value of `a`—except at time 0, when `a_lag` is 0 for all subjects. Similarly, `l_lag` is the lagged version of `l`. The variable `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 outcome `y` is coded as missing at the first visit; this is because it is assumed that all subjects are event-free at the beginning of follow-up. Even if `y` were given a value of 1 for some subjects at visit 0, this would be ignored and the subject would be treated as event-free (and hence lost to follow-up between visits 0 and 1). To avoid this, such subjects should be dropped from the dataset before using `gformula`.

The treatment and confounder variables, as well as any variables derived from them, are coded as missing at the final time point. This is because the outcome measured at the final visit refers to whether the event took place in the time between the penultimate and final visits, and hence cannot be affected by the treatment or confounder values *at* the final visit; thus, these treatment/confounder values are irrelevant to the problem and are not included.

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

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

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

$$\frac{\exp\left(3 - L_{t-1} - 0.3\sum_{s=0}^{t-1} A_s - U\right)}{1 + \exp\left(3 - L_{t-1} - 0.3\sum_{s=0}^{t-1} A_s - U\right)}$$

  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

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

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

$$\frac{\exp\left(3 - L_9 - 0.3\sum_{s=0}^{9} A_s - U\right)}{1 + \exp\left(3 - L_9 - 0.3\sum_{s=0}^{9} A_s - U\right)}$$

The second dataset is generated in the same way except that there is censoring, due to both death and 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

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

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

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

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):

```
. use tvc2
. list id t d y l a cuma a_lag cuma_lag l_lag, sepby(id)
```

id	t	d	y	l	a	cuma	a_lag	cuma_lag	l_lag

(*output omitted*)

4	0	.	.	6.016037	1	1	0	0	0
4	1	0	0	6.55031	0	1	1	1	6.016037
4	2	0	0	5.914584	0	1	0	1	6.55031
4	3	0	0	5.277947	0	1	0	1	5.914584
4	4	0	0	4.789021	1	2	0	1	5.277947
4	5	0	0	5.451276	1	3	1	2	4.789021
4	6	0	0	5.740773	1	4	1	3	5.451276
4	7	0	0	6.124776	1	5	1	4	5.740773
4	8	0	0	6.563818	0	5	1	5	6.124776

(*output omitted*)

9	0	.	.	5.392478	0	0	0	0	0
9	1	0	0	4.794044	1	1	0	0	5.392478
9	2	1	.	.	.	.	1	1	4.794044

(*output omitted*)

12	0	.	.	5.29574	1	1	0	0	0
12	1	0	0	5.711647	0	1	1	1	5.29574
12	2	0	0	5.272157	1	2	0	1	5.711647
12	3	0	0	5.691969	0	2	1	2	5.272157
12	4	0	0	5.160069	0	2	0	2	5.691969
12	5	0	1	4.758551	0	2	0	2	5.160069
13	0	.	.	5.801145	1	1	0	0	0
13	1	0	0	6.207112	0	1	1	1	5.801145
13	2	0	0	5.646282	0	1	0	1	6.207112
13	3	0	0	5.058614	0	1	0	1	5.646282
13	4	0	0	4.745497	1	2	0	1	5.058614
13	5	0	0	5.517207	1	3	1	2	4.745497
13	6	0	0	5.979877	1	4	1	3	5.517207
13	7	0	0	6.352118	0	4	1	4	5.979877
13	8	0	0	5.699408	0	4	0	4	6.352118
13	9	0	0	5.128589	0	4	0	4	5.699408
13	10	0	0	.	.	.	0	4	5.128589

(*output omitted*)

Subject 9 died between visits 1 and 2. Subject 13 remained AIDS-free to the end of follow-up. Subject 12 developed AIDS between visits 4 and 5. Subject 4 was lost to follow-up after visit 8.

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, outcome(y)          ///
 commands(y:logit, l:regress, a:logit)                             ///
 equations(y:l_lag cuma_lag, l:l_lag a_lag, a:l a_lag)             ///
 idvar(id) tvar(t) varyingcovariates(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 laggedvars(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. $(A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9) = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$
2. $(A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9) = (0, 0, 1, 1, 1, 1, 1, 1, 1, 1)$
3. $(A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9) = (0, 0, 0, 0, 1, 1, 1, 1, 1, 1)$
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. $(A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9) = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$

These regimes are defined only through A_9 (in the code, this is ensured by the inclusion of the condition `t<10` in the definition of each regime) because A_{10} is irrelevant to the outcome, as discussed in the section *The Data* above.

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, outcome(y)              ///
 commands(y:logit, l:regress, a:logit)                                 ///
 equations(y:l_lag cuma_lag, l:l_lag a_lag, a:l a_lag)                 ///
 idvar(id) tvar(t) varyingcovariates(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 laggedvars(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 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. Alternatively, in this setting, one might want to compare dynamic regimes of the form "start treatment the first time L falls beneath x and then continue treatment". For example, if $x = 6.9$, this would be coded as

```
a=0 if t<10 & l>6.9 & a_lag=0 \ a=1 if t<10 & (l<=6.9 | a_lag=1)
```

Finally, we analyze the second dataset (with losses to follow-up and censoring due to death), and we 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, outcome(y)               ///
 commands(y:logit, l:regress, a:logit, d:logit)                           ///
 equations(y:l_lag cuma_lag, l:l_lag a_lag, a:l a_lag, d:l_lag cuma_lag)  ///
 idvar(id) tvar(t) varyingcovariates(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                     ///
 laggedvars(l_lag a_lag cuma_lag) msm(stcox cuma_lag) derived(cuma)       ///
 lagrules(l_lag: l 1, a_lag: a 1, cuma_lag: cuma 1)                       ///
 derrules(cuma:cuma_lag+a) death(d) seed(79)
```

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
```

y	G-computation estimate of Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
c1	-.4620501	.0426871	-10.82	0.000	-.5457153	-.3783849

```
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
```

y	G-computation estimate of av. log IR	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
Int. 1	-3.710399	.1178156	-31.49	0.000	-3.941313	-3.479485
Int. 2	-2.849232	.0737148	-38.65	0.000	-2.99371	-2.704754
Int. 3	-2.409732	.0742438	-32.46	0.000	-2.555247	-2.264216
Int. 4	-2.155157	.0708308	-30.43	0.000	-2.293983	-2.016331
Int. 5	-1.992489	.0690772	-28.84	0.000	-2.127878	-1.8571
Int. 6	-2.010118	.0656089	-30.64	0.000	-2.138709	-1.881526
Obs. regime						
simulated	-2.693125	.0648117	-41.55	0.000	-2.820153	-2.566096
observed	-2.585342					

```
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.

        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
```

y	G-computation estimate of cum. incidence	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
Int. 1	.208	.0217588	9.56	0.000	.1653535	.2506465
Int. 2	.408	.0211903	19.25	0.000	.3664678	.4495322
Int. 3	.565	.0242743	23.28	0.000	.5174232	.6125768
Int. 4	.677	.0251431	26.93	0.000	.6277205	.7262795
Int. 5	.77	.0256334	30.04	0.000	.7197594	.8202406
Int. 6	.782	.0248577	31.46	0.000	.7332798	.8307202
Obs. regime						
simulated	.486	.0222683	21.82	0.000	.4423549	.5296451
observed	.519					

All three tables point toward a beneficial effect of treatment: the more treatment a subject receives, the longer the subject survives AIDS-free. This is seen from the negative log hazard ratio associated with cumulative treatment (corresponding to a hazard ratio of 0.630, 95% CI $[0.579, 0.685]$) 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. Of the study participants, 78% were simulated as having developed AIDS during the hypothetical study in which treatment was withheld (intervention 6), whereas only 21% were simulated to have developed AIDS when treatment was prescribed at all times (intervention 1).

There is a small difference between the simulated and observed data under the observational regime (49% versus 52% for the cumulative incidences and −2.69 versus −2.59 for the average log incidence rates). These differences are small relative to the standard error.

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
```

y	G-computation estimate of av. log IR	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
Int. 1	-3.751415	.1141952	-32.85	0.000	-3.975233	-3.527596
Int. 2	-3.568099	.107682	-33.14	0.000	-3.779152	-3.357046
Int. 3	-3.506201	.1037585	-33.79	0.000	-3.709564	-3.302838
Int. 4	-2.795603	.0673776	-41.49	0.000	-2.927661	-2.663546
Int. 5	-2.322542	.0605097	-38.38	0.000	-2.441139	-2.203945
Obs. regime simulated	-2.719811	.0675269	-40.28	0.000	-2.852161	-2.587461
observed	-2.585342					

```
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.

         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
```

y	G-computation estimate of cum. incidence	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
Int. 1	.203	.0215237	9.43	0.000	.1608144	.2451856
Int. 2	.236	.0213897	11.03	0.000	.1940769	.2779231
Int. 3	.252	.022857	11.03	0.000	.2072012	.2967988
Int. 4	.451	.0222252	20.29	0.000	.4074395	.4945605
Int. 5	.635	.0224428	28.29	0.000	.5910129	.6789871
Obs. regime simulated	.479	.0235225	20.36	0.000	.4328967	.5251033
observed	.519					

We are not able to estimate the parameters of an MSM from this analysis, as explained above. However, the results from the average log incidence rates and the cumulative incidences confirm that treatment is beneficial. Higher AIDS-free survival is achieved under the dynamic regime in which x, the threshold below which ART is administered, is highest. The observational regime appears to lie between regime 4 and regime 5 in terms of AIDS-free survival.

The results from this sort of analysis can be combined with information on the cost of treatment in a cost–benefit analysis to determine the optimal regime.

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
```

y	G-computation estimate of Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
c1	-.3664798	.0439505	-8.34	0.000	-.4526213	-.2803384

```
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
```

y	G-computation estimate of av. log IR	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
Int. 1	-3.711201	.1217193	-30.49	0.000	-3.949767	-3.472636
Int. 2	-2.877049	.0811772	-35.44	0.000	-3.036153	-2.717944
Int. 3	-2.426325	.0757291	-32.04	0.000	-2.574751	-2.277899
Int. 4	-2.265722	.0784065	-28.9	0.000	-2.419396	-2.112048
Int. 5	-2.210928	.0755939	-29.25	0.000	-2.35909	-2.062767
Int. 6	-2.13881	.0735494	-29.08	0.000	-2.282964	-1.994656
Obs. regime						
simulated	-2.886523	.0731307	-39.47	0.000	-3.029857	-2.74319
observed	-2.876808					

```
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.

        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
```

y	G-computation estimate of cum. incidence	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
Int. 1 (o)	.206	.022829	9.02	0.000	.161256	.250744
(d)	.024	.0062109	3.86	0.000	.0118269	.0361731
Int. 2 (o)	.382	.0221304	17.26	0.000	.3386251	.4253749
(d)	.078	.0110279	7.07	0.000	.0563857	.0996143
Int. 3 (o)	.517	.0243066	21.27	0.000	.46936	.56464
(d)	.102	.0160495	6.36	0.000	.0705437	.1334563
Int. 4 (o)	.583	.0277243	21.03	0.000	.5286613	.6373387
(d)	.122	.0187849	6.49	0.000	.0851822	.1588178
Int. 5 (o)	.612	.0290661	21.06	0.000	.5550314	.6689686
(d)	.149	.0217222	6.86	0.000	.1064252	.1915748
Int. 6 (o)	.656	.0293983	22.31	0.000	.5983804	.7136196
(d)	.149	.0225586	6.61	0.000	.104786	.193214
Obs. regime						
simulated (o)	.405	.0242117	16.73	0.000	.357546	.452454
(d)	.069	.0117705	5.86	0.000	.0459303	.0920697
observed (o)	.4					
(d)	.068					
(l)	.201					

Key: (o) = outcome, (d) = death, (l) = lost to follow-up

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

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 or loss to follow-up).

```
. use tvc1, clear
. stset t, id(id) failure(y)
  (output omitted)
. stcox cuma_lag
  (output omitted)
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cuma_lag	.7040005	.0375827	-6.57	0.000	.6340625	.7816528

```
. stcox cuma_lag l
  (output omitted)
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cuma_lag	1.376655	.0991092	4.44	0.000	1.195486	1.585278
l	.2933428	.0246312	-14.61	0.000	.2488298	.3458187

The unadjusted analysis suggests a slightly less beneficial effect of treatment than we found from the g-computation analysis. This is what we would expect because the unadjusted analysis fails to take into account that the treated subjects at any given visit are less healthy than the untreated subjects (because the decision of whether to treat depends on CD4 count at that visit). Adjusting for log(CD4) makes things even worse and suggests that treatment is harmful, because conditioning on future CD4 count masks most of the beneficial effect of the treatment, and collider-stratification bias is induced, exaggerating the bias further.

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

```
. use tvc2, clear
. stset t, id(id) failure(y)
  (output omitted)
. stcox cuma_lag
  (output omitted)
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cuma_lag	.6669321	.0409318	-6.60	0.000	.5913446	.7521814

```
. stcox cuma_lag l
  (output omitted)
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cuma_lag	1.372486	.1198532	3.63	0.000	1.156581	1.628694
l	.2893885	.0297036	-12.08	0.000	.2366529	.3538757

4.2 Example II: Mediation

The data

A dataset comprising 10,000 subjects was simulated according to the description given in section 1.2 as follows:

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

Independently and randomly, 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 only GGT missing, 424 have only BMI missing, and 427 have only alcohol 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`.

```
. use mediation
. list sep alc bmi log_ggt sbp log_ggt_alc alc_sbp log_ggt_sbp in 1/3
```

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

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 outcome(sbp) equations(bmi:i.sep alc, log_ggt:bmi alc, sbp:bmi alc   ///
 log_ggt log_ggt_alc i.sep) commands(bmi:regress, log_ggt:regress, sbp:regress) ///
 exposure(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)
```

The output

Here is the (abridged) output:

	G-computation estimate	Bootstrap Std. Err.	z	P>|z|	Normal-based [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 SBP. If everyone were to stop drinking, then 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; that is, it acts through BMI and other pathways.

Comparison with standard analysis

Here are the standard analyses that 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. Five proper imputations for each missing value are drawn (using `ice` in Stata), and the results are analyzed and combined using the `mim` command. Such multiple proper imputations are now required because 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)
[note: using ice-style format variables _mi and _mj]
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	-6.7303	-5.67264	0.015
_Isep_3	-11.1104	.336718	-33.00	0.000	-11.7706	-10.4503	0.022
alc	3.27134	.066665	49.07	0.000	3.14041	3.40226	0.082
_cons	123.994	.262807	471.81	0.000	123.479	124.509	0.023

```
. xi: mim: regress sbp i.sep alc log_ggt bmi
i.sep             _Isep_1-3           (naturally coded; _Isep_1 omitted)
[note: using ice-style format variables _mi and _mj]
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	-5.96281	-4.97551	0.043
_Isep_3	-10.3605	.311652	-33.24	0.000	-10.9718	-9.74911	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	4.62397	5.0315	0.044
bmi	.435676	.051427	8.47	0.000	.334853	.536499	0.023
_cons	99.0651	1.28187	77.28	0.000	96.5522	101.578	0.016

These estimates are not directly comparable with estimates obtained using the g-computation procedure because the coefficient of alcohol in the analyses above are for a unit change in units consumed per day. Because 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 $3.27 \times 2.2 = 7.25$, which 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,

because some of the direct effect of alcohol on SBP (that is, 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 as the number of time points increases. In the time-varying confounding example above, with $T = 9$, fitting the parametric models, simulating the data under each intervention, and then analyzing each simulated dataset takes around 30 seconds on a standard PC. Thus, if 1,000 bootstrap samples are required, then 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 allows 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 has been measured. In addition, the procedure requires that correct parametric models be postulated for the postbaseline variables in the observational data. For the estimation of NDEs and NIEs, further assumptions are required, as discussed in section 2.2.

Alternative semiparametric models and estimation methods have been proposed by Robins et al. (1992), Robins (1999), and Robins, Hernán, and Brumback (2000). These involve g-estimation of structural nested models and inverse probability weighted estimation of MSMs. These alternative methods rely on fewer parametric modeling 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 by Sterne and Tilling (2002) and Fewell et al. (2004).

In parallel with these developments in epidemiological literature, alternative methods based on instrumental variables, balancing scores, and event-history analysis have been proposed in econometrics literature (Lechner and Miquel 2001; Lechner 2001; Miquel 2002, 2003; Abbring 2003; Lechner 2004).

However, the g-computation procedure has some clear advantages over alternative methods: it can more easily deal with complex multivariate (or joint) interventions, and it can easily compare a wide range of static and dynamic regimes, as well as the observational regime, which can be important in informing policy (Taubman et al. 2009). These advantages are in addition to the increased statistical efficiency that is gained at the price of stronger modeling assumptions (Daniel et al. 2011).

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 because 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 make the g-computation procedure more accessible to a wider audience of applied researchers.

6 Acknowledgments

`gformula.ado` uses the `detangle` and `formatlist` procedures from `ice.ado`, with kind permission from Patrick Royston.

The command and this article have benefited from very useful comments and suggestions from Daniela Zugna, Deborah Ford, Linda Harrison, and an anonymous referee. We thank all of them for their time, patience, and interest.

7 References

Abbring, J. H. 2003. Dynamic econometric program evaluation. Discussion Paper No. 804, Institute for the Study of Labor (IZA). http://ftp.iza.org/dp804.pdf.

Angrist, J. D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Cain, L. E., J. M. Robins, E. Lanoy, R. Logan, D. Costagliola, and M. A. Hernán. 2010. When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *International Journal of Biostatistics* 6(2): Article 18.

D'Agostino, R. B., M.-L. Lee, A. J. Belanger, L. A. Cupples, K. Anderson, and W. B. Kannel. 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., S. N. Cousens, B. L. De Stavola, M. G. Kenward, and J. A. C. Sterne. 2011. Methods for dealing with time-varying confounding. Unpublished manuscript.

Daniel, R. M., B. L. De Stavola, and S. N. Cousens. Forthcoming. Time-varying confounding: Some practical considerations in a likelihood framework. In *Causality: Statistical Perspectives and Applications*, ed. C. Berzuini, A. P. Dawid, and L. Bernardinelli. Wiley.

Didelez, V., A. P. Dawid, and S. Geneletti. 2006. Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 138–146. Arlington, VA: AUAI Press.

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

Gill, R. D., and J. M. Robins. 2001. Causal inference for complex longitudinal data: The continuous case. *Annals of Statistics* 29: 1785–1811.

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

Hafeman, D. M. 2009. "Proportion explained": A causal interpretation for standard measures of indirect effect? *American Journal of Epidemiology* 170: 1443–1448.

Hafeman, D. M., and T. J. VanderWeele. 2011. Alternative assumptions for the identification of direct and indirect effects. *Epidemiology* 22: 753–764.

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

Hubbard, A. E., and M. J. van der Laan. 2008. Population intervention models in causal inference. *Biometrika* 95: 35–47.

Lechner, M. 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies*, ed. M. Lechner and F. Pfeiffer, 43–58. Heidelberg: Physica-Verlag.

———. 2004. Sequential matching estimation of dynamic causal models. Discussion Paper No. 1042, Institute for the Study of Labor (IZA). http://ftp.iza.org/dp1042.pdf.

Lechner, M., and R. Miquel. 2001. A potential outcome approach to dynamic programme evaluation: Nonparametric identification. Discussion Paper No. 2001-07, Department of Economics, University of St. Gallen.
http://www.alexandria.unisg.ch/export/DL/33987.pdf.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: Wiley.

Miquel, R. 2002. Identification of dynamic treatments effects by instrumental variables. Discussion Paper No. 2002-11, University of St. Gallen.

———. 2003. Identification of effects of dynamic treatments with a difference-in-differences approach. Discussion Paper No. 2003-06, University of St. Gallen.
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=388660.

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

Murphy, S. A. 2003. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B* 65: 331–366.

Pearl, J. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference in Uncertainty in Artificial Intelligence*, 411–420. San Francisco: Morgan Kaufmann.

———. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York: Cambridge University Press.

Petersen, M. L., S. E. Sinisi, and M. J. van der Laan. 2006. Estimation of direct causal effects. *Epidemiology* 17: 276–284.

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.

———. 1987a. Addendum to "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers and Mathematics with Applications* 14: 923–945.

———. 1987b. Errata to "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers and Mathematics with Applications* 14: 917–921.

———. 1989a. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Services Research Methodology: A Focus on AIDS*, ed. L. Sechrest, H. Freeman, and A. Mulley. Washington, DC: U.S. Public Health Service.

———. 1989b. Errata to "Addendum to 'A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect'". *Computers and Mathematics with Applications* 18: 477.

———. 1997. Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, ed. M. Berkane, 69–117. New York: Springer.

———. 1999. Association, causation, and marginal structural models. *Synthese* 121: 151–179.

———. 2003. Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems*, ed. P. Green, N. L. Hjort, and S. Richardson, 70–81. New York: Oxford University Press.

Robins, J. M., D. Blevins, G. Ritter, and M. Wulfsohn. 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., and R. D. Gill. 1997. Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine* 16: 39–56.

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

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

Robins, J. M., and M. A. Hernán. 2009. Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis*, ed. G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, 553–599. Boca Raton, FL: Chapman & Hall/CRC.

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

Robins, J. M., and T. S. Richardson. 2010. Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, ed. P. Shrout. Oxford: Oxford University Press.

Rothman, K. J., S. Greenland, and T. L. Lash. 2008. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins.

Sterne, J. A. C., and K. Tilling. 2002. G-estimation of causal effects, allowing for time-varying confounding. *Stata Journal* 2: 164–182.

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

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

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

About the authors

Rhian Daniel is a research fellow at the London School of Hygiene and Tropical Medicine. Bianca De Stavola is a professor of biostatistics, and Simon Cousens is a professor of epidemiology and medical statistics, both at the London School of Hygiene and Tropical Medicine. The authors worked together on this article and the `gformula` command as part of a grant entitled *Quantitative methods for the assessment of systematic error in observational studies: Improving causal research*, funded by the Medical Research Council, UK (Grant number: G0701024).